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 general form, and with the theory of the discriminant, done now over a field F. Then we discuss the Jordan form and its basic applications to physics, and other advanced decomposition results. We then go into positivity topics, involving matrices, bilinear forms, with a look at curved space-time, and discrete Laplacians. Finally, we discuss the various groups of matrices, of reflection type, of Lie type, of classical and quantum physics type, and of arithmetic type.

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 instead on the underlying geometry. Personally I tend to regard any vector as something alive and dynamic, and any linear map as 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, determinant and diagonalization, that you surely know over $\mathbb{R}$ and $\mathbb{C}$, by doing this over a field F. 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, 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 the form of groups of matrices. Nothing or almost can resist to these frightening formations, and we will 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 arithmetic groups, 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 [9]. 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 official plan is, 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.

Contents

Preface 3

Part I. Linear algebra 9

Chapter 1. Linear maps 11
- 1a. Linear maps 11
- 1b. Real geometry 18
- 1c. Complex numbers 22
- 1d. Arbitrary fields 25
- 1e. Exercises 28

Chapter 2. Matrix theory 29
- 2a. Matrix inversion 29
- 2b. The determinant 35
- 2c. Diagonalization 44
- 2d. Field closures 48
- 2e. Exercises 48

Chapter 3. Spectral theorems 49
- 3a. Self-adjoints 49
- 3b. Rotations, unitaries 53
- 3c. Normal matrices 54
- 3d. Polar decomposition 57
- 3e. Exercises 58

Chapter 4. Polynomials, roots 59
- 4a. Resultant 59
- 4b. Discriminant 63
- 4c. Low dimensions 66
- 4d. Density tricks 72
- 4e. Exercises 74

Part II. Advanced results 75

Chapter 5. Jordan form 77
5a. Abstract algebra 77
5b. Jordan form 77
5c. Basic examples 77
5d. Spectral measures 77
5e. Exercises 77

Chapter 6. Dynamical systems 79
6a. Differential equations 79
6b. Matrix exponential 79
6c. Complex functions 79
6d. Some applications 79
6e. Exercises 79

Chapter 7. Singular values 81
7a. Triangularization 81
7b. Decomposition results 81
7c. Singular values 81
7d. Some applications 81
7e. Exercises 81

Chapter 8. Compact operators 83
8a. Infinite matrices 83
8b. Compact operators 88
8c. Singular values 92
8d. Elliptic operators 96
8e. Exercises 96

Part III. Positive matrices 97

Chapter 9. Hessian matrices 99
9a. Calculus, Jacobian 99
9b. Second derivatives 99
9c. Positive matrices 99
9d. Higher derivatives 99
9e. Exercises 99

Chapter 10. Forms, signature 101
10a. Bilinear forms 101
10b. Riemannian manifolds 101
10c. Curved spacetime 101
10d. Lorentz geometry 101
10e. Exercises 101

Chapter 11. Bistochastic matrices 103
11a. Circulant matrices 103
11b. Fourier, Hadamard 103
11c. Bistochastic matrices 103
11d. Sinkhorn algorithm 103
11e. Exercises 103

Chapter 12. Graphs and designs 105
12a. Discrete Laplacian 105
12b. Into the waves 105
12c. Into the heat 105
12d. Design theory 105
12e. Exercises 105

Part IV. Geometric aspects 107

Chapter 13. Finite groups 109
13a. Matrix groups 109
13b. Abelian groups 109
13c. Peter-Weyl 109
13d. Reflection groups 109
13e. Exercises 109

Chapter 14. Lie theory 111
14a. Exponential, revised 111
14b. Lie algebras 111
14c. Cases ABCD 111
14d. Cases EFG 111
14e. Exercises 111

Chapter 15. Spin matrices 113

15a. Pauli matrices 113
15b. Euler-Rodrigues 113
15c. Dirac matrices 113
15d. Clifford and Weyl 113
15e. Exercises 113

Chapter 16. Arithmetic groups 115
16a. General theory 115
16b. Semisimplicity 115
16c. Into arithmetic 115
16d. Absolute groups 115
16e. Exercises 115

Bibliography 117

Part I

Linear algebra

You're my heart
You're my soul
I keep it shining
Everywhere I go

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 [75], or Lax-Terrell [67].

(2) Then comes $\mathbb{R}^2$. This is the entry point into 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 vulgar 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 [76], and Lax-Terrell [68].

(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 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, and 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 not know. 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 develop the general theory over an arbitrary field 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) *And by replacing, whenever possible, $\mathbb{R}$ by 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 [63] if you are more into algebra, and Lax [63] if you are more into analysis. You can check also my basic linear algebra book [9], 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 arbitrary fields F for a bit 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, the equivalence with (1) is something clear, and 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 on $\mathbb{R}x$ must be a certain multiple of x, and we are led in this way to the following formula:

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

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

$$\begin{aligned}(P_x)_{ij} &= < P_x e_j, e_i > \\ &= \frac{1}{||x||^2} < e_j, x >< x, e_i > \\ &= \frac{x_j x_i}{||x||^2}\end{aligned}$$

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

With the above done, 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. Indeed, this classification is best done by first proceeding modulo linear transformations, and then in general. We have:

THEOREM 1.10. *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. This is best done by using first linear transformations, as follows:

(1) Let us first classify the conics up to non-degenerate linear transformations of the plane, which are by transformations as follows, assumed to be invertible:

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

Our claim is that as solutions we have the circles, parabolas, hyperbolas, along with some degenerate solutions, namely $\emptyset$, points, lines, pairs of lines, $\mathbb{R}^2$.

(2) 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.

(3) Getting now to the proof of our claim in (1), 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 $l = 0$.

(4) 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 $f = 0$.

(5) Thus, done with the classification, up to linear transformations as in (1). But this classification leads to the classification in general too, by applying now linear transformations to the solutions that we found, with the conclusions in the statement. □

Of course, we can use such methods for other geometric problems, and we will be back to this, later in this book, when discussing more in detail manifolds and geometry.

Finally, no discussion about matrices and linear algebra could be complete without a word on multivariable calculus, and the matrices appearing there. We first have:

THEOREM 1.11. *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. □

We have as well a matrix at order 2, for the scalar functions, as follows:

THEOREM 1.12. *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

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

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

PROOF. This is something more 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 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. □

1c. Complex numbers

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 1.13. *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:

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

We have as well a second eigenvector, as follows:

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

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 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 $x_1 + \ldots + x_N = 0$. In the complex setting, however, the roots of unity come to the rescue, via:

PROPOSITION 1.14. *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. □

We have the following definition, inspired by what happens in Proposition 1.14:

DEFINITION 1.15. *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 1.16. *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. According to the discussion in Proposition 1.5, we are left with finding the 0-eigenvectors of the all-one matrix, which amounts in solving the following equation:

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

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 1.14:

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

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 several 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 1.17. *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 an inverse $a^{-1} \in F$, satisfying $aa^{-1} = 1$.*
(3) *The sum and product are compatible via $a(b + c) = ab + ac$.*

As basic examples of fields, we have of course $\mathbb{Q}, \mathbb{R}, \mathbb{C}$. Another basic example is $\mathbb{F}_p$, the integers modulo a prime number p. More generally, associated to any prime power $q = p^k$ is a certain field $\mathbb{F}_q$ having q elements, and these are the only finite fields:

THEOREM 1.18. *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 a first observation, associated to any finite field F is its characteristic, the smallest number $p \in \mathbb{N}$ such that the following equality happens, inside F:

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

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. □

As a comment here, the characteristic p formula appearing in the above proof, namely $1+\ldots+1=0$, with p terms in the sum, can potentially cause troubles with linear algebra and geometry over $\mathbb{F}_q$. We will see in a moment illustrations for this.

In relation with prime numbers and arithmetic, we have as well the following construction, which is more subtle, producing a characteristic 0 field, as we like them:

THEOREM 1.19. *Given p prime and $a/b\in\mathbb{Q}$, we can write $a/b=p^k(c/d)$ with c,d prime to p, and set $|a/b|=p^{-k}$. The function $|\,.\,|\to\mathbb{Q}$, called p-adic absolute value, or even p-adic norm, is not exactly a norm, but $d(x,y)=|x-y|$ is a distance on $\mathbb{Q}$, 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 standard, with the completion procedure in the statement being 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.

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 1.20. *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:

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

Also, 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. □

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 1.21. *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 1.22. *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. □

As a conclusion, better stay away from characteristic p. And this is what we will do in this book, unless in the last chapter, 16, dealing with specialized arithmetic aspects.

1e. Exercises

Exercises:

Exercise 1.23.

Exercise 1.24.

Exercise 1.25.

Exercise 1.26.

Exercise 1.27.

Exercise 1.28.

Exercise 1.29.

Exercise 1.30.

Bonus exercise.

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:

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

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 from the above that, in order for A to be invertible, the 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. 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,*

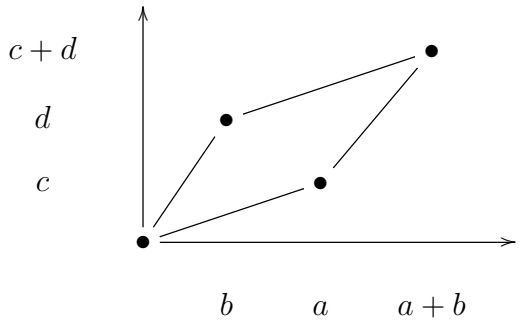
$$\det^+ \begin{pmatrix} a & b \\ c & d \end{pmatrix} = |ad - bc|$$

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

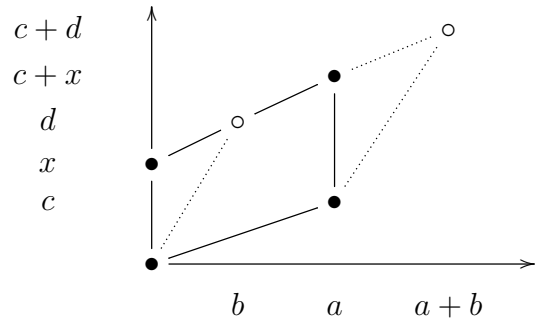
PROOF. We must show that the area of the parallelogram formed by $\binom{a}{c}, \binom{b}{d}$ equals $|ad - bc|$. We can assume $a, b, c, d > 0$ for simplifying, the proof in general being similar. Moreover, by switching if needed the vectors $\binom{a}{c}, \binom{b}{d}$, we can assume that we have:

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

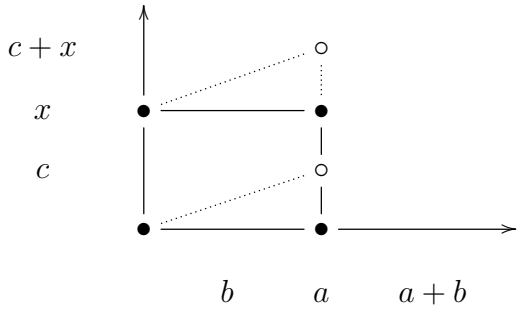
According to these conventions, the picture of our parallelogram is as follows:



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:



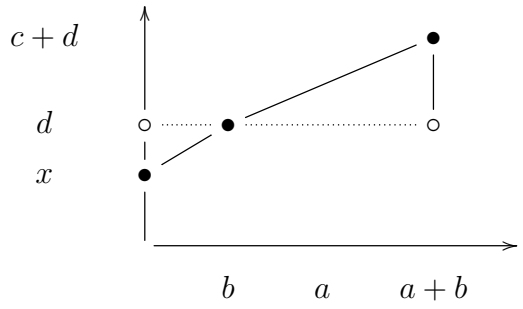
We can further modify this parallelogram, once again by not altering its area, by sliding the right side downwards, until we reach the Ox axis:



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:

$$ax=ad-bc$$

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, call them 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, 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 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 from the fact that the formula in the statement produces a certain function $\det : M_N(\mathbb{R}) \to \mathbb{R}$, which has the 4 properties in 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 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, 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, and in order to get familiar with it, there is nothing better than doing some computations. As a first, basic example, 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. We can indeed recover this formula as well as a particular case of Theorem 2.25. To be more precise, the permutations in the statement are listed according to the

lexicographic order, and the computation of the corresponding signatures is something elementary, by using the various rules from Theorem 2.24. □

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, and we will be back to this, on several occassions. But, most importantly, that formula allows us to deal now with the complex matrices too, by formulating the following statement:

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 has been done 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 section, in relation with volumes, which don't extend well to $\mathbb{C}^N$. □

Good news, this is the end of the general theory that we wanted to develop. We have now in our bag all the needed techniques for computing the determinant.

2c. 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 2.29. *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 2.30. *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 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 characteristic polynomial, which appears via the following fundamental result:

THEOREM 2.31. *Given a matrix $A \in M_N(\mathbb{C})$, consider its characteristic polynomial:*

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

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

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

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

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

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

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

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

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

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

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

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 2.32. *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 2.33. *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 2.31, 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 2.30 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 2.33. Let us record as well a useful algorithmic version of the above result:

THEOREM 2.34. *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 2.33, 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.

2d. Field closures

Field closures.

2e. Exercises

Exercises:

Exercise 2.35.

Exercise 2.36.

Exercise 2.37.

Exercise 2.38.

Exercise 2.39.

Exercise 2.40.

Exercise 2.41.

Exercise 2.42.

Bonus exercise.

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:

PROPOSITION 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:

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

We therefore have $\lambda \in \{0, 1\}$, as claimed, and as a final conclusion here, the diagonalization of the self-adjoint matrices is 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, $P^2 = P = P^t$, 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 12.3, 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:

$$\begin{aligned} < Ax, x > &= < CC^*x, x > \\ &= < C^*x, C^*x > \\ &\geq 0 \end{aligned}$$

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

$$\begin{aligned} < Ax, x > &= < x, A^*x > \\ &= < A^*x, x > \end{aligned}$$

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

$$\begin{aligned} < Ax, x > &= < \lambda x, x > \\ &= \lambda < x, x > \\ &\geq 0 \end{aligned}$$

(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 its square is given by:

$$\begin{aligned} B^2 &= U\sqrt{D}U^* \cdot U\sqrt{D}U^* \\ &= UDU^* \\ &= A \end{aligned}$$

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

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. □

3b. Rotations, unitaries

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

THEOREM 3.7. *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 we obtain $\lambda \in \mathbb{T}$, as claimed. Our next claim now is that the eigenspaces corresponding to different eigenvalues are pairwise orthogonal. Assume indeed that:

$$Uv = \lambda v \quad , \quad Uw = \mu w$$

We have then the following computation, using $U^* = U^{-1}$ and $\lambda, \mu \in \mathbb{T}$:

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

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 3.8. *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 3.7. □

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 simplest unitaries that we know, namely the rotations in the real plane, we have the following formula, that we know well from chapter 1:

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

Many other things can be said about the diagonalization of rotations, and we will be back to this later in this book, when 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 3.9. *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 finish. Indeed, we can decompose the problem, and the matrix A itself, following these eigenspaces of 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 3.7. □

3d. Polar decomposition

As a first application of all this, we have the following result:

THEOREM 3.10. *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 3.9, 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. □

We can now formulate a first polar decomposition result, as follows:

THEOREM 3.11. *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. This is routine, and follows by comparing the actions of $A, |A|$ on the vectors $v \in \mathbb{C}^N$, and deducing from this the existence of a unitary $U \in U_N$ as above. □

Observe that at $N = 1$ we obtain in this way the usual polar decomposition of the nonzero complex numbers. More generally now, we have the following result:

THEOREM 3.12. *Any square 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. Again, this follows by comparing the actions of $A, |A|$ on the vectors $v \in \mathbb{C}^N$, and deducing from this the existence of a partial isometry U as above. Alternatively, we can get this from Theorem 3.11, applied on the complement of the 0-eigenvectors. □

3e. Exercises

Exercises:

EXERCISE 3.13.

EXERCISE 3.14.

EXERCISE 3.15.

EXERCISE 3.16.

EXERCISE 3.17.

EXERCISE 3.18.

EXERCISE 3.19.

EXERCISE 3.20.

Bonus exercise.

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 $P \in F[X]$. 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,

$$R(P, Q) = \begin{vmatrix} p_k & & & q_l & & \\ \vdots & \ddots & & \vdots & \ddots & \\ p_0 & & p_k & q_0 & & q_k \\ & \ddots & \vdots & & \ddots & \vdots \\ & & p_0 & & & q_0 \end{vmatrix}$$

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. □

4c. Low dimensions

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, coming from Theorem 4.1, but let us see how this comes via the formula in Theorem 4.6, namely:

$$\Delta(P) = a^{2N-2} \prod_{i<j}(r_i - r_j)^2$$

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:

$$\begin{aligned}\Delta(P) &= a^4(r - z)^2(r - \bar{z})^2(z - \bar{z})^2\\ &= a^4|r - z|^4(2iIm(z))^2\\ &= -4a^4|r - z|^4Im(z)^2\\ &< 0\end{aligned}$$

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

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 $108=4\times27$. 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. □

There are many more things that can be said, along these lines, and with the remark however that, once you get to $\mathbb{C}$, the Cardano formula can become unuseful.

In higher degree things become complicated. In degree 4, to start with, we first have the following result, dealing with the discriminant and its applications:

THEOREM 4.10. *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 4.10:

PROPOSITION 4.11. *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 4.10. □

In degree 5 and more, things become fairly complicated, and the conceptual explanations for what happens here come from the Galois theory of field extensions.

However, in higher degree we do have some arithmetic tricks, for computing the integer or rational roots of the polynomials having integer of rational coefficients.

There are a lot of analytic tricks too, in relation with the general question of computing zeros of polynomials, starting with Newton's old method, in the real case. Also, in the complex case, we have countless tricks, coming from complex analysis.

4d. Density tricks

Getting back now to linear algebra, we can use the above resultant and discriminant technology, in relation with our diagonalization questions, as follows:

THEOREM 4.12. *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. □

We can now formulate a quite tricky and powerful result, as follows:

THEOREM 4.13. *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 surface $\det A = 0$, 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 surface 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 4.12. 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 4.14. *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 4.13, 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 4.13 (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.

(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 4.13 (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.

Finally, let us 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.

4e. Exercises

Exercises:

Exercise 4.15.

Exercise 4.16.

Exercise 4.17.

Exercise 4.18.

Exercise 4.19.

Exercise 4.20.

Exercise 4.21.

Exercise 4.22.

Bonus exercise.

Part II

Advanced results

Cheri, Cheri lady
Going through emotion
Love is where you find it
Listen to your heart

CHAPTER 5

Jordan form

5a. Abstract algebra

We must first do some abstract algebra, for the eigenspaces.

5b. Jordan form

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

5c. Basic examples

Basic examples. As an application of this, we can recover the density of the diagonalizable matrices, that we can get via the Jordan form, by perturbing the diagonal.

5d. Spectral measures

Any normal matrix has a spectral measure, formed by the Dirac masses at the eigenvalues, but in the non-normal case, things can be quite complicated. In particular, we have some interesting computations here for the Jordan blocks.

5e. Exercises

Exercises:

EXERCISE 5.1.

EXERCISE 5.2.

EXERCISE 5.3.

EXERCISE 5.4.

EXERCISE 5.5.

EXERCISE 5.6.

EXERCISE 5.7.

EXERCISE 5.8.

Bonus exercise.

CHAPTER 6

Dynamical systems

6a. Differential equations

Differential equations, and their basic theory.

6b. Matrix exponential

We are led in this way into computing matrix exponentials.

6c. Complex functions

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 via 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.

6d. Some applications

Back to dynamical systems and ordinary differential equations, with some applications, based on the Jordan form, and on the functional calculus results developed above.

6e. Exercises

Exercises:

EXERCISE 6.1.

EXERCISE 6.2.

EXERCISE 6.3.

EXERCISE 6.4.

EXERCISE 6.5.

EXERCISE 6.6.

EXERCISE 6.7.

Exercise 6.8.

Bonus exercise.

CHAPTER 7

Singular values

7a. Triangularization

Triangularization.

7b. Decomposition results

Decomposition results.

7c. Singular values

Singular values.

7d. Some applications

Some applications.

7e. Exercises

Exercises:

Exercise 7.1.

Exercise 7.2.

Exercise 7.3.

Exercise 7.4.

Exercise 7.5.

Exercise 7.6.

Exercise 7.7.

Exercise 7.8.

Bonus exercise.

CHAPTER 8

Compact operators

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*

$$l^2(I) = \left\{ (x_i)_{i \in I} \Big| \sum_i |x_i|^2 < \infty \right\}$$

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 abstract, we have, more generally, the following result, which shows that our formalism covers as well the Schrödinger spaces of type $L^2(\mathbb{R}^3)$:

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 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. Compact operators

Let us start with a basic definition, as follows:

DEFINITION 8.7. *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 8.8. *The set of finite rank operators*

$$F(H) \subset B(H)$$

is a two-sided $*$-ideal.

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

(1) It is clear from definitions that $F(H)$ is indeed a vector space, with this due to the following formulae, valid for any $S, T \in B(H)$, which are both clear:

$$\dim(Im(S+T)) \leq \dim(Im(S)) + \dim(Im(T))$$

$$\dim(Im(\lambda T)) = \dim(Im(T))$$

(2) Let us prove now that $F(H)$ is stable under $*$. Given $T \in F(H)$, we can regard it as an invertible operator between finite dimensional Hilbert spaces, as follows:

$$T : (\ker T)^\perp \to Im(T)$$

Thus, 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:

$$\begin{aligned} \dim(Im(T^*)) &= \dim(\overline{Im(T^*)}) \\ &= \dim((\ker T)^\perp) \\ &= \dim(Im(T)) \end{aligned}$$

(3) Finally, regarding 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 8.9. *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 8.10. *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 8.11. *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 from definitions that $K(H)$ is indeed a vector space, with this due to the following formulae, valid for any $S, T \in B(H)$, which are both clear:

$$(S+T)(B_1) \subset S(B_1) + T(B_1)$$

$$(\lambda T)(B_1) = |\lambda| \cdot T(B_1)$$

(2) In order to prove now that $K(H)$ is closed, assume that a sequence $T_n \in K(H)$ converges to $T \in B(H)$. Given $\varepsilon > 0$, let us 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, with $S \subset B_1$ being finite:

$$T(B_1) \subset \bigcup_{x \in S} B_{3\varepsilon}(Tx)$$

But this shows that our limiting operator T is compact, as desired.

(3) Regarding the fact that $K(H)$ is stable under involution, this follows from Proposition 8.8, Proposition 8.10 and (2). Indeed, by using Proposition 8.10, 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 obtain that we have as well:

$$T^* = \lim_{n \to \infty} T_n^*$$

We know from Proposition 8.8 that the operators T_n^* are of finite rank, and so compact by Proposition 8.10, and by using (2) we obtain that T^* is compact too, as desired.

(4) Finally, regarding 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:

$$(ST)(B_1) = S(T(B_1))$$

$$(TS)(B_1) \subset ||S|| \cdot T(B_1)$$

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

Here is now a second key result regarding the compact operators:

THEOREM 8.12. *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 was assumed to be compact, and the sequence $\{e_n\}$ is bounded, a certain subsequence $\{Te_{n_k}\}$ must converge. Thus, by replacing once again $\{e_n\}$ with a subsequence, we can assume that the following holds, with $x \neq 0$:

$$Te_n \to x$$

But this is a contradiction, because we obtain in this way:

$$\begin{aligned} < x, x > &= \lim_{n\to\infty} < Te_n, x > \\ &= \lim_{n\to\infty} < e_n, T^* x > \\ &= 0 \end{aligned}$$

Thus our assumption $Te_n \not\to 0$ was wrong, and we obtain the result.

"$\impliedby$" Assume $Te_n \to 0$, for any orthonormal system $\{e_n\} \subset H$. In order to prove that T is compact, we use the various results established above, which show that this is the same as proving 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 the above is wrong, and so that there exists $\varepsilon > 0$ such that the following holds:

$$S \in F(H) \implies ||T - S|| > \varepsilon$$

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 such vectors $e_1, \ldots, e_n$. Let $E \subset H$ be the linear space spanned by these vectors, and let us set:

$$P = Proj(E)$$

Since the operator TP has finite rank, our assumption above shows that we have:

$$||T - TP|| > \varepsilon$$

Thus, we can find a vector $x \in H$ 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. □

8c. Singular values

Let us discuss now the spectral theory of the compact operators. We first have:

PROPOSITION 8.13. *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 8.14. *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 8.15. *If T is compact and self-adjoint, there is an orthogonal basis of H made of eigenvectors of T.*

PROOF. We use Proposition 8.13. 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 8.14 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 8.16. *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 8.13 (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 8.15, 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 8.13 (2).

(4) This is something that we know too, from Proposition 8.15. □

Finally, we have the following result, regarding the general case:

THEOREM 8.17. *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 8.16, 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 8.16 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. □

8d. Elliptic operators

Elliptic operators.

8e. Exercises

Exercises:

Exercise 8.18.

Exercise 8.19.

Exercise 8.20.

Exercise 8.21.

Exercise 8.22.

Exercise 8.23.

Exercise 8.24.

Exercise 8.25.

Bonus exercise.

Part III

Positive matrices

Geronimo's Cadillac
It's tossing oh in your head
It's tossing and turning
It's burning, it makes you mad

CHAPTER 9

Hessian matrices

9a. Calculus, Jacobian

Calculus, Jacobian.

9b. Second derivatives

Second derivatives.

9c. Positive matrices

Positive matrices.

9d. Higher derivatives

Higher derivatives.

9e. Exercises

Exercises:

Exercise 9.1.

Exercise 9.2.

Exercise 9.3.

Exercise 9.4.

Exercise 9.5.

Exercise 9.6.

Exercise 9.7.

Exercise 9.8.

Bonus exercise.

CHAPTER 10

Forms, signature

10a. Bilinear forms

Bilinear forms.

10b. Riemannian manifolds

Riemannian manifolds.

10c. Curved spacetime

Curved spacetime.

10d. Lorentz geometry

Lorentz geometry.

10e. Exercises

Exercises:

Exercise 10.1.

Exercise 10.2.

Exercise 10.3.

Exercise 10.4.

Exercise 10.5.

Exercise 10.6.

Exercise 10.7.

Exercise 10.8.

Bonus exercise.

CHAPTER 11

Bistochastic matrices

11a. Circulant matrices

Circulant matrices.

11b. Fourier, Hadamard

Fourier, Hadamard.

11c. Bistochastic matrices

Bistochastic matrices.

11d. Sinkhorn algorithm

Sinkhorn algorithm.

11e. Exercises

Exercises:

Exercise 11.1.

Exercise 11.2.

Exercise 11.3.

Exercise 11.4.

Exercise 11.5.

Exercise 11.6.

Exercise 11.7.

Exercise 11.8.

Bonus exercise.

CHAPTER 12

Graphs and designs

12a. Discrete Laplacian

Discrete Laplacian.

12b. Into the waves

Into the waves.

12c. Into the heat

Into the heat.

12d. Design theory

Design theory.

12e. Exercises

Exercises:

Exercise 12.1.

Exercise 12.2.

Exercise 12.3.

Exercise 12.4.

Exercise 12.5.

Exercise 12.6.

Exercise 12.7.

Exercise 12.8.

Bonus exercise.

Part IV

Geometric aspects

You can win if you want
If you want it, you will win
On your way, you will see
That life is more than fantasy

CHAPTER 13

Finite groups

13a. Matrix groups

Matrix groups.

13b. Abelian groups

Abelian groups.

13c. Peter-Weyl

Peter-Weyl.

13d. Reflection groups

Reflection groups.

13e. Exercises

Exercises:

EXERCISE 13.1.

EXERCISE 13.2.

EXERCISE 13.3.

EXERCISE 13.4.

EXERCISE 13.5.

EXERCISE 13.6.

EXERCISE 13.7.

EXERCISE 13.8.

Bonus exercise.

CHAPTER 14

Lie theory

14a. Exponential, revised

Exponential, revised.

14b. Lie algebras

Lie algebras.

14c. Cases ABCD

Cases ABCD.

14d. Cases EFG

Cases EFG.

14e. Exercises

Exercises:

Exercise 14.1.

Exercise 14.2.

Exercise 14.3.

Exercise 14.4.

Exercise 14.5.

Exercise 14.6.

Exercise 14.7.

Exercise 14.8.

Bonus exercise.

CHAPTER 15

Spin matrices

15a. Pauli matrices

Pauli matrices.

15b. Euler-Rodrigues

Euler-Rodrigues.

15c. Dirac matrices

Dirac matrices.

15d. Clifford and Weyl

Clifford and Weyl.

15e. Exercises

Exercises:

Exercise 15.1.

Exercise 15.2.

Exercise 15.3.

Exercise 15.4.

Exercise 15.5.

Exercise 15.6.

Exercise 15.7.

Exercise 15.8.

Bonus exercise.

CHAPTER 16

Arithmetic groups

16a. General theory

General theory.

16b. Semisimplicity

Semisimplicity.

16c. Into arithmetic

Into arithmetic.

16d. Absolute groups

Absolute groups.

16e. Exercises

Congratulations for having read this book, and no exercises for this final chapter.

Bibliography

[1] V.I. Arnold, Ordinary differential equations, Springer (1973).

[2] V.I. Arnold, Mathematical methods of classical mechanics, Springer (1974).

[3] V.I. Arnold, Lectures on partial differential equations, Springer (1997).

[4] V.I. Arnold, Catastrophe theory, Springer (1974).

[5] V.I. Arnold and B.A. Khesin, Topological methods in hydrodynamics, Springer (1998).

[6] M.F. Atiyah, K-theory, CRC Press (1964).

[7] M.F. Atiyah, The geometry and physics of knots, Cambridge Univ. Press (1990).

[8] M.F. Atiyah and I.G. MacDonald, Introduction to commutative algebra, Addison-Wesley (1969).

[9] T. Banica, Linear algebra and group theory (2023).

[10] T. Banica, Graphs and their symmetries (2024).

[11] T. Banica, Basic number theory (2024).

[12] R.J. Baxter, Exactly solved models in statistical mechanics, Academic Press (1982).

[13] I. Bengtsson and K. Życzkowski, Geometry of quantum states, Cambridge Univ. Press (2006).

[14] N. Berline, E. Getzler and M. Vergne, Heat kernels and Dirac operators, Springer (2004).

[15] B. Blackadar, K-theory for operator algebras, Cambridge Univ. Press (1986).

[16] S.J. Blundell and K.M. Blundell, Concepts in thermal physics, Oxford Univ. Press (2006).

[17] S.M. Carroll, Spacetime and geometry, Cambridge Univ. Press (2004).

[18] A. Connes, Noncommutative geometry, Academic Press (1994).

[19] A. Connes and M. Marcolli, Noncommutative geometry, quantum fields and motives, AMS (2008).

[20] W.N. Cottingham and D.A. Greenwood, An introduction to the standard model of particle physics, Cambridge Univ. Press (2012).

[21] H.S.M. Coxeter, Regular polytopes, Dover (1948).

[22] W. de Launey and D. Flannery, Algebraic design theory, AMS (2011).

[23] P.A.M. Dirac, Principles of quantum mechanics, Oxford Univ. Press (1930).

[24] M.P. do Carmo, Differential geometry of curves and surfaces, Dover (1976).

[25] M.P. do Carmo, Riemannian geometry, Birkhäuser (1992).

[26] S. Dodelson, Modern cosmology, Academic Press (2003).

[27] S.K. Donaldson, Riemann surfaces, Oxford Univ. Press (2004).

[28] R. Durrett, Probability: theory and examples, Cambridge Univ. Press (1990).

[29] A. Einstein, Relativity: the special and the general theory, Dover (1916).

[30] L.C. Evans, Partial differential equations, AMS (1998).

[31] W. Feller, An introduction to probability theory and its applications, Wiley (1950).

[32] E. Fermi, Thermodynamics, Dover (1937).

[33] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics I: mainly mechanics, radiation and heat, Caltech (1963).

[34] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics II: mainly electromagnetism and matter, Caltech (1964).

[35] R.P. Feynman, R.B. Leighton and M. Sands, The Feynman lectures on physics III: quantum mechanics, Caltech (1966).

[36] R.P. Feynman and A.R. Hibbs, Quantum mechanics and path integrals, Dover (1965).

[37] P. Flajolet and R. Sedgewick, Analytic combinatorics, Cambridge Univ. Press (2009).

[38] A.P. French, Special relativity, Taylor and Francis (1968).

[39] W. Fulton, Algebraic topology, Springer (1995).

[40] W. Fulton and J. Harris, Representation theory, Springer (1991).

[41] C. Godsil and G. Royle, Algebraic graph theory, Springer (2001).

[42] H. Goldstein, C. Safko and J. Poole, Classical mechanics, Addison-Wesley (1980).

[43] D.J. Griffiths, Introduction to electrodynamics, Cambridge Univ. Press (2017).

[44] D.J. Griffiths and D.F. Schroeter, Introduction to quantum mechanics, Cambridge Univ. Press (2018).

[45] D.J. Griffiths, Introduction to elementary particles, Wiley (2020).

[46] J. Harris, Algebraic geometry, Springer (1992).

[47] R. Hartshorne, Algebraic geometry, Springer (1977).

[48] M.P. Hobson, G.P. Efstathiou and A.N. Lasenby, General relativity, Cambridge Univ. Press (2006).

[49] L. Hörmander, The analysis of linear partial differential operators, Springer (1983).

[50] R.A. Horn and C.R. Johnson, Matrix analysis, Cambridge Univ. Press (1985).

[51] K. Huang, Introduction to statistical physics, CRC Press (2001).

[52] K. Huang, Fundamental forces of nature, World Scientific (2007).

[53] J.E. Humphreys, Introduction to Lie algebras and representation theory, Springer (1972).

[54] N. Jacobson, Basic algebra, Dover (1974).

[55] V.F.R. Jones, Subfactors and knots, AMS (1991).

[56] V.F.R. Jones and V.S Sunder, Introduction to subfactors, Cambridge Univ. Press (1997).

[57] T. Kibble and F.H. Berkshire, Classical mechanics, Imperial College Press (1966).

[58] M. Kumar, Quantum: Einstein, Bohr, and the great debate about the nature of reality, Norton (2009).

[59] T. Lancaster and K.M. Blundell, Quantum field theory for the gifted amateur, Oxford Univ. Press (2014).

[60] L.D. Landau and E.M. Lifshitz, Mechanics, Pergamon Press (1960).

[61] L.D. Landau and E.M. Lifshitz, The classical theory of fields, Addison-Wesley (1951).

[62] L.D. Landau and E.M. Lifshitz, Quantum mechanics: non-relativistic theory, Pergamon Press (1959).

[63] S. Lang, Algebra, Addison-Wesley (1993).

[64] S. Lang, Abelian varieties, Dover (1959).

[65] P. Lax, Linear algebra and its applications, Wiley (2007).

[66] P. Lax, Functional analysis, Wiley (2002).

[67] P. Lax and M.S. Terrell, Calculus with applications, Springer (2013).

[68] P. Lax and M.S. Terrell, Multivariable calculus with applications, Springer (2018).

[69] J.M. Lee, Introduction to topological manifolds, Springer (2011).

[70] J.M. Lee, Introduction to smooth manifolds, Springer (2012).

[71] J.M. Lee, Introduction to Riemannian manifolds, Springer (2019).

[72] M.L. Mehta, Random matrices, Elsevier (2004).

[73] M.A. Nielsen and I.L. Chuang, Quantum computation and quantum information, Cambridge Univ. Press (2000).

[74] B.M. Peterson and B. Ryden, Foundations of astrophysics, Cambridge Univ. Press (2010).

[75] W. Rudin, Principles of mathematical analysis, McGraw-Hill (1964).

[76] W. Rudin, Real and complex analysis, McGraw-Hill (1966).

[77] W. Rudin, Fourier analysis on groups, Dover (1972).

[78] B. Ryden, Introduction to cosmology, Cambridge Univ. Press (2002).

[79] D.V. Schroeder, An introduction to thermal physics, Oxford Univ. Press (1999).

[80] B. Schutz, A first course in general relativity, Cambridge Univ. Press (2009).

[81] J.P. Serre, A course in arithmetic, Springer (1973).

[82] J.P. Serre, Linear representations of finite groups, Springer (1977).

[83] I.R. Shafarevich, Basic algebraic geometry, Springer (1974).

[84] J.H. Silverman, The arithmetic of elliptic curves, Springer (1986).

[85] J.H. Silverman and J.T. Tate, Rational points on elliptic curves, Springer (2015).

[86] B. Singh, Basic commutative algebra, World Scientific (2011).

[87] A.M. Steane, Thermodynamics, Oxford Univ. Press (2016).

[88] A.M. Steane, Relativity made relatively easy, Oxford Univ. Press (2012).

[89] D.R. Stinson, Combinatorial designs: constructions and analysis, Springer (2006).

[90] C.H. Taubes, Differential geometry, Oxford Univ. Press (2011).

[91] J.R. Taylor, Classical mechanics, Univ. Science Books (2003).

[92] D.V. Voiculescu, K.J. Dykema and A. Nica, Free random variables, AMS (1992).

[93] J. von Neumann, Mathematical foundations of quantum mechanics, Princeton Univ. Press (1955).

[94] S. Weinberg, Foundations of modern physics, Cambridge Univ. Press (2011).

[95] S. Weinberg, Lectures on quantum mechanics, Cambridge Univ. Press (2012).

[96] S. Weinberg, Lectures on astrophysics, Cambridge Univ. Press (2019).

[97] S. Weinberg, Cosmology, Oxford Univ. Press (2008).

[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).