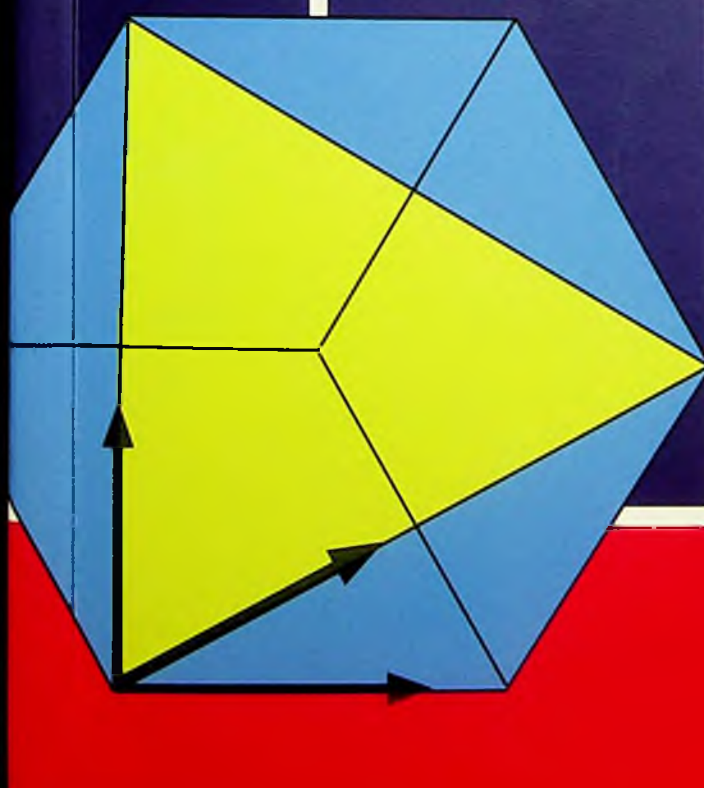
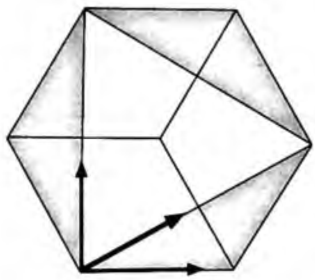
LINEAR ALGEBRA

Examples and Applications



Alain M. Robert



LINEAR ALGEBRA

Examples and Applications

Alain M. Robert
Université de Neuchâtel, Switzerland

World Scientific

NEW JERSEY • LONDON • SINGAPORE • SHANGHAI • HONG KONG • TAIPEI • BANGALORE

Published by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

LINEAR ALGEBRA
Examples and Applications

ISBN 981-256-432-2
ISBN 981-256-499-3 (pbk)

Printed in Singapore by B & JO Enterprise

To the memory of my parents,
who understood that
I would never make a good farmer

Foreword

Mathematics is a living creation, and linear algebra has undergone a real metamorphosis during the twentieth century, partly due to the birth and development of computers. It is so active that entire periodical magazines are now devoted to it, and one single book can only reflect part of its vitality. Here is an attempt to face this challenge in a concise—although rigorous—manner. Linear algebra is a general and powerful language. This book is *based on examples and applications*, justifying the elaboration of such an abstract language.

In the ***first part***, vector spaces are approached through carefully chosen linear systems, and linear maps are introduced through matrix multiplication. The four initial chapters constitute the skeleton of the linear category. The importance and ubiquity of this structure is emphasized by the applications of the rank theory (Chapter 5), and in the geometric approach to eigenvectors (Chapter 6). Since even and odd functions appear as the eigenspaces of the symmetry operator, we do not assume *a priori* finite dimensionality, and bases are discussed and examples are given in the general context.

The ***second part*** is devoted to the study of metric relations (angles, orthogonality) in real vector spaces. Several geometric properties can easily be derived from an inner product. The best approximation theorem, with its application to the mean squares method is certainly the most used in practice. Here bilinearity appears on the scene, and this fascinating property culminates in the abstract form of duality (Chapter 9).

Finally, the ***third part*** is rooted in volume computations, revealing the phenomenon of multi-linearity. Hence, determinants come last (Chapter 10), and constitute the golden adornment of the theory. They play an essential part in the algebraic properties of eigenvalues. The main result proved in this book is the spectral theorem (for real symmetric matrices in Chapter 8, and for normal operators in Chapter 12). Its geometrical meaning is emphasized with the polar decomposition for linear maps between finite-dimensional real vector spaces.

A few appendices contain independent complements. Of special importance is the appendix on finite probability spaces, where the notion of independence for random variables is compared with that of linear independence.

As is probably apparent, this book is written for curious and motivated students in physics, chemistry, computer science, engineering,... and not solely for mathematicians. I believe that our duty is to form scientists capable of understanding each other's problems. Having in the same class students interested in various disciplines provides an opportunity to show them the relevance of mathematics through linear algebra, by selecting examples that might catch their interest. It should not be wasted on teaching them to perform mechanical manipulations based on a set of axioms, a task better suited to a computer! This is why I have tried to minimize the axiomatic aspect, leaving out the discussion of general fields, assuming implicitly that the scalars are real (or complex) numbers. But I have chosen general proofs of the main theorems (and in particular for the rank theorem), relegating the use of inner products and orthogonality (specific to real numbers) to the second part of the book, as already mentioned.

These students are supposed to have a previous acquaintance with basic calculus and to be familiar with the language of arrows for maps, their composition, and inversion. Only a brief summary of set theory is included. Another prerequisite concerns vectors in two and three dimensions, Cartesian and polar coordinates (elementary trigonometry). Hence this text is directed to students who follow (or have previously followed) a first calculus course. This is particularly apparent with the examples concerning polynomials and their derivatives, linear fractional transformations, and rational functions.

Needless to say, exercises, tutorials (or individual support in any form) are essential to check that the students understand and can apply this theory. Since books with many routine exercises are easily available, I have limited the number of such exercises. On the other hand, more difficult problems have been included (with hints, or even complete solutions).

If I have tried to bring the main facts to the forefront, I have made no effort to satisfy all the needs of future research mathematicians, or theoretical physicists. They will have to complete this study by examining vector spaces over any field (or even modules over principal ideal domains) and tensor products.

I have chosen to avoid the discussion of the normal Jordan form. In my opinion, its importance is best revealed with a specific application in mind: Markov chain theory, coupled linear differential systems, Riesz theory for compact operators in Banach spaces, linear algebraic groups (where additive and multiplicative Jordan decompositions both appear); each provides such an opportunity. My purpose was only to convey the basic aspects of this cornerstone in mathematical education.

Finally, I have to thank O. Besson, A. Gertsch Hamadene, and A. Junod who read parts of preliminary versions of this book, detected several mistakes and made useful suggestions.

April 2005

Alain M. Robert

Contents

Chapter 1

Linear Systems: Elimination Method

A principal objective of linear algebra is the resolution of systems of linear equations (no product of unknown variables occurs: A precise definition will be given later). We present this topic by example, starting from the high school point of view, assuming that two by two and three by three systems have already been considered.

1.1 Examples of Linear Systems

1.1.1 A Review Example

Suppose that three unknown numbers x, y, and z are linked by the relations

$$y + z = 1, \quad z + x = 2, \quad x + y = 3.$$

Are there any (or many) possibilities for these numbers x, y, z? How can we find them? The answer to this problem consists in solving the system of three equations

$$\begin{cases} y + z = 1 \\ z + x = 2 \\ x + y = 3, \end{cases}$$

in three *variables*. Notice that we also consider that the first equation, in which x does not appear explicitly, concerns the three unknown variables x, y, and z: In fact, we can say that the coefficient of x in this equation is 0 (zero). To discuss this system, we are going to transform it into simpler ones, *having the same solutions*. First of all, we rewrite it in a more conventional way, letting the variables appear in alphabetical order in each equation

$$\begin{cases} \phantom{x+{}} y + z = 1 \\ x \phantom{{}+y} + z = 2 \\ x + y \phantom{{}+z} = 3. \end{cases}$$

It is better to start with an equation containing the first variable x, so let us exchange the first two equations (the well chosen right-hand sides emphasize this operation) :

$$\begin{cases} x \phantom{{}+y} + z = 2 \\ \phantom{x+{}} y + z = 1 \\ x + y \phantom{{}+z} = 3. \end{cases}$$

Now, we eliminate the variable x in the last two equations: For this purpose, we subtract the first one from the last one

$$\begin{cases} x \phantom{{}+y} + z = 2 \\ \phantom{x+{}} y + z = 1 \\ \phantom{x+{}} y - z = 1. \end{cases}$$

In this way, the last two equations concern the variables y, z only. Let us subtract the second equation from the third one

$$\begin{cases} x \phantom{{}+y} + z = 2 \\ \phantom{x+{}} y + z = 1 \\ - 2z = 0. \end{cases}$$

The last equation does not contain the variable y any more: It requires $2z = 0$, hence $z = 0$. The second equation informs us now that $y = 1$. Finally, the first equation leads to $x = 2$. The *solution set* is the list

$$\begin{cases} x = 2 \\ y = 1 \\ z = 0. \end{cases}$$

1.1.2 Covering a Sphere with Hexagons and Pentagons

Question to a bee:

Is it possible to cover the surface of a sphere with hexagons only?

Answer by a mathematician:

No, it is impossible!

How can one show that nobody will be able to do it, if each of our attempts fails? One method consists in replacing the question by a more general one, where there are some possibilities, and in fact where all possibilities have a common feature not realized by hexagons only.

Let us try to cover the surface of a sphere with (curved) hexagons *and* pentagons. By convention, we juxtapose two polygons along a common edge, three polygons having a common vertex. Such configurations occur in biology, chemistry, architecture, sport,... It is easy to find a few equations (or relations) linking the unknown numbers of such polygons. Let us introduce

$$x : \text{number of pentagons}, \quad y : \text{number of hexagons},$$
$$e : \text{number of edges}, \quad f : \text{number of faces}, \quad v : \text{number of vertices}.$$

The number of faces is equal to the sum of the number of pentagons and the number of hexagons, hence a first obvious relation: $f = x + y$ (hence the introduction of the variable f could be avoided, replacing it systematically by $x+y$; but since we are aiming at a general method, valid for large systems, this extra variable adds interest to the example). Since each pentagon has five edges, and each hexagon has six, the expression $5x + 6y$ counts twice the number of edges (each edge belongs to exactly two polygons). Hence a second relation

$$5x + 6y = 2e.$$

Similarly, since each vertex belongs to three polygons, the sum $5x + 6y$ also counts vertices three times (by convention, we are assuming that three polygons only meet at each vertex), and we get

$$5x + 6y = 3v.$$

From this follows $2e = 3v$, but this relation tells us nothing new since it is a consequence of the previous ones. Another, more subtle relation has been discovered by Euler

$$f + v = e + 2$$

(we indicate a proof in the Appendix to this section). We have obtained a system consisting of four equations linking the five variables x, y, e, f, and v:

$$\begin{cases} x+y &= f \\ 5x+6y &= 2e \\ 5x+6y &= 3v \\ f+v &= e+2. \end{cases}$$

Let us rewrite these equations, grouping the variables in the left-hand side

$$\begin{cases} f-x-y &= 0 \\ 2e-5x-6y &= 0 \\ 3v-5x-6y &= 0 \\ e-f-v &= -2. \end{cases}$$

As with the previous worked-out example, we are going to transform this system into simpler, equivalent ones (having the same solutions). This tedious procedure will be simplified if we only write the coefficients of the equations, adopting the order e, f, v, x, y for the unknown variables. Hence instead of the first equation

$$f - x - y = 0,$$

which represents the relation

$$0e + 1f + 0v - 1x - 1y = 0$$

in these five variables, we simply write the row of its coefficients

$$0 \quad 1 \quad 0 \quad -1 \quad -1 \quad | \quad 0.$$

The separator " | " distinguishes the left-hand from the right-hand sides. Such an abbreviation is only meaningful if we write a 0 (zero coefficient) for variables not explicitly present in the equation, and keep in mind the chosen order of the variables, namely here

$$1st = e, \quad 2nd = f, \quad 3rd = v, \quad 4th = x, \quad 5th = y.$$

This row notation keeps track of the correct position of the variables. With a similar row notation for the other three equations, the system is now abbreviated by a *rectangular array* containing four rows

$$\left(\begin{array}{ccccc|c} 0 & 1 & 0 & -1 & -1 & 0 \\ 2 & 0 & 0 & -5 & -6 & 0 \\ 0 & 0 & 3 & -5 & -6 & 0 \\ 1 & -1 & -1 & 0 & 0 & -2. \end{array}\right.$$

We can now start transforming this system into simpler, equivalent ones. It is advisable to start the system by an equation containing the first variable. So we exchange the first and last equations and obtain an equivalent system

$$\left(\begin{array}{ccccc|c} 1 & -1 & -1 & 0 & 0 & -2 \\ 2 & 0 & 0 & -5 & -6 & 0 \\ 0 & 1 & 0 & -1 & -1 & 0 \\ 0 & 0 & 3 & -5 & -6 & 0 \end{array}\right).$$

The big parentheses are only used to isolate the system from the context. As with the first worked-out example, we try to get rid of the first variable in the second, third, and fourth equations, so that they only concern the four remaining variables f, v, x, and y. For this purpose, let us subtract twice the first equation from the second one

$$\left(\begin{array}{ccccc|c} 1 & -1 & -1 & 0 & 0 & -2 \\ 0 & 2 & 2 & -5 & -6 & 4 \\ 0 & 1 & 0 & -1 & -1 & 0 \\ 0 & 0 & 3 & -5 & -6 & 0 \end{array}\right).$$

It is essential to observe that this new system has the same solutions as the previous one, simply since we may add twice the first equation to the new second one, and recover the previous one. If we permute the two central equations

$$\begin{pmatrix} 1 & -1 & -1 & 0 & 0 & | & -2 \\ 0 & 1 & 0 & -1 & -1 & | & 0 \\ 0 & 2 & 2 & -5 & -6 & | & 4 \\ 0 & 0 & 3 & -5 & -6 & | & 0 \end{pmatrix},$$

the coefficient of the variable f in the second equation is 1, and can be used to get rid of the second variable from the third equation on. Hence, from the third equation, we subtract twice the second, obtaining

$$\begin{pmatrix} 1 & -1 & -1 & 0 & 0 & | & -2 \\ 0 & 1 & 0 & -1 & -1 & | & 0 \\ 0 & 0 & 2 & -3 & -4 & | & 4 \\ 0 & 0 & 3 & -5 & -6 & | & 0 \end{pmatrix}.$$

Here, the last two equations concern only v, x, and y. If we multiply the third equation by $\frac{1}{2}$, its leading coefficient is transformed into a 1:

$$\begin{pmatrix} 1 & -1 & -1 & 0 & 0 & | & -2 \\ 0 & 1 & 0 & -1 & -1 & | & 0 \\ 0 & 0 & 1 & -3/2 & -2 & | & 2 \\ 0 & 0 & 3 & -5 & -6 & | & 0 \end{pmatrix}.$$

To eliminate v from the last equation, we may subtract from it the triple of the preceding one:

$$\begin{pmatrix} 1 & -1 & -1 & 0 & 0 & | & -2 \\ 0 & 1 & 0 & -1 & -1 & | & 0 \\ 0 & 0 & 1 & -3/2 & -2 & | & 2 \\ 0 & 0 & 0 & -5+9/2 & -6+6 & | & -6 \end{pmatrix}.$$

We have now reached the system

$$\begin{pmatrix} 1 & -1 & -1 & 0 & 0 & | & -2 \\ 0 & 1 & 0 & -1 & -1 & | & 0 \\ 0 & 0 & 1 & -3/2 & -2 & | & 2 \\ 0 & 0 & 0 & -1/2 & 0 & | & -6 \end{pmatrix},$$

having a last row corresponding to the equation $-x/2 = -6$, namely

$$x = 12.$$

Here comes a *surprise*: Although the system has fewer equations than variables, the value of x is uniquely determined

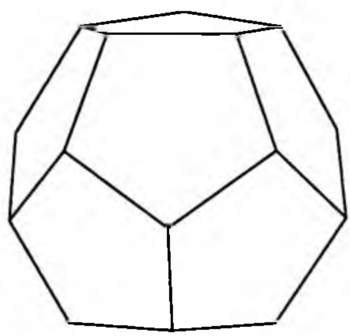
> In any subdivision of the sphere consisting in hexagons and pentagons only, *the number of pentagons is fixed and equal to* 12.

Isn't this remarkable! On the other hand, several examples will now show that the number of hexagons is not fixed.

(a) A partition of the sphere is easily obtained with twelve pentagons and no hexagon:

$$\begin{cases} x = 12 \\ y = 0. \end{cases}$$

Simply consider a regular dodecahedron inscribed in the sphere

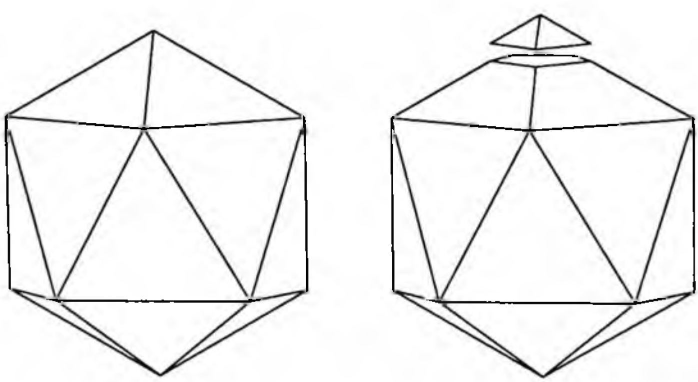


and project its twelve pentagonal faces on the surface of the sphere.

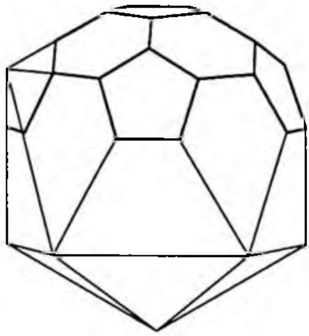
(b) Another solution

$$\begin{cases} x = 12 \\ y = 20. \end{cases}$$

is also obtained as follows. Start with a regular icosahedron (12 vertices and 20 faces formed by equilateral triangles). Cut the vertices, replacing them by pentagonal faces, as in the following picture.



When this is repeated at each vertex, the triangular faces are replaced by hexagons.



Eventually, one obtains a polyhedron having $12 \times 5 = 60$ vertices. These vertices give the positions of the carbon atoms in the buckminsterfullerene C_{60}.

(c) One can construct a geometrical solution with $y = 2$ as follows. Start with six pentagons attached to one hexagon. This roughly covers a hemisphere. Two such hemispheres—placed symmetrically—will cover the sphere.

Comment. Notice that many purely algebraic solutions of the system have no geometrical realization. For example, one may take $y = \frac{1}{2}$ and adapt correspondingly

$$e = 31.5, \quad f = 12.5, \quad v = 21 \quad (\text{and } x = 12).$$

Similarly, one can take $y = -1$ together with

$$e = 27, \quad f = 11, \quad v = 18 \quad (\text{and } x = 12).$$

More generally, one can take y arbitrarily, say $y = t$, together with

$$e = 3t + 30, \quad f = t + 12, \quad v = 2t + 20 \quad (\text{and } x = 12).$$

This is the *general solution* of the system. It depends on the choice of a *parameter* t. Also notice that one could decide to choose e arbitrarily, and deduce expressions for the other variables y, f, and v (but still $x = 12$). The problem of determining which solutions of the linear system in five variables do have a geometric realization is a difficult one (not tackled by linear algebra). An obvious necessary condition is that y should be a nonnegative integer. But this condition is not sufficient.

1.1.3 A Literal Example

From my own experience, the elimination method looks deceptively simple and it is necessary to practice it on several examples.

Somebody might be looking for a solution of the following linear system in the variables x, y, z, and u:

$$\begin{cases} x+y+z+8u &= 6 \\ x+y+8z+u &= 1 \\ x+8y+z+u &= 2 \\ 8x+y+z+u &= 0. \end{cases}$$

Having afterthoughts, he might prefer solutions of

$$\begin{cases} x+y+z+7u &= 6.5 \\ x+y+7z+u &= 1.1 \\ x+7y+z+u &= 2 \\ 7x+y+z+u &= 0. \end{cases}$$

And so on... This is a good reason for considering a more general system from the outset, having literal coefficients

$$(S) \qquad \begin{cases} x+y+z+au &= A \\ x+y+az+u &= B \\ x+ay+z+u &= C \\ ax+y+z+u &= D. \end{cases}$$

Here the letters a, A, B, C, and D represent known values, or *parameters*, on which the solution(s) will depend.

As before, we write the rows of coefficients instead of the equations, and represent the whole system by a rectangular array:

$$\left(\begin{array}{cccc|c} 1 & 1 & 1 & a & A \\ 1 & 1 & a & 1 & B \\ 1 & a & 1 & 1 & C \\ a & 1 & 1 & 1 & D \end{array}\right).$$

Keeping the first row fixed, we subtract multiples of it from the other ones, in order to eliminate the first variable in the next rows. Here is what we obtain

$$\left(\begin{array}{cccc|c} 1 & 1 & 1 & a & A \\ 0 & 0 & a-1 & 1-a & B-A \\ 0 & a-1 & 0 & 1-a & C-A \\ 0 & 1-a & 1-a & 1-a^2 & D-aA \end{array}\right).$$

(a) When $a=1$, we have

$$\left(\begin{array}{cccc|c} 1 & 1 & 1 & 1 & A \\ 0 & 0 & 0 & 0 & B-A \\ 0 & 0 & 0 & 0 & C-A \\ 0 & 0 & 0 & 0 & D-A \end{array}\right),$$

and there is only one nontrivial equation: The first one. The last three equations (having only 0's in front of the separator) lead to *compatibility conditions*

$$\begin{cases} 0 &= B - A \\ 0 &= C - A \\ 0 &= D - A. \end{cases}$$

Hence the system is *consistent* only when

$$A = B = C = D.$$

(*b*) When $a \neq 1$, we permute the second and third rows, in order to bring a nonzero coefficient (of y) in the second place (of the the second row)

$$\left(\begin{array}{cccc|c} 1 & 1 & 1 & a & A \\ 0 & a-1 & 0 & 1-a & C-A \\ 0 & 0 & a-1 & 1-a & B-A \\ 0 & 1-a & 1-a & 1-a^2 & D-aA \end{array}\right).$$

If we add the second row to the last one, we eliminate the second variable from the third row on. Hence we achieve a column of zeros under this crucial coefficient, called *second pivot* (a precise definition follows)

$$\left(\begin{array}{cccc|c} 1 & 1 & 1 & a & A \\ 0 & a-1 & 0 & 1-a & C-A \\ 0 & 0 & a-1 & 1-a & B-A \\ 0 & 0 & 1-a & 2-a-a^2 & D-aA+C-A \end{array}\right).$$

Notice that the last column keeps track of the operations made, and in particular shows how to reverse them to come back to the initial system. To place a zero under the *third pivot*, we still add the third row to the last one

$$\left(\begin{array}{cccc|c} 1 & 1 & 1 & a & A \\ 0 & a-1 & 0 & 1-a & C-A \\ 0 & 0 & a-1 & 1-a & B-A \\ 0 & 0 & 0 & 3-2a-a^2 & D-aA+C-A+B-A \end{array}\right).$$

If $a^2 + 2a - 3 = 0$, the last row leads to a compatibility condition. The roots of this quadratic equation are $a = -3$ and $a = 1$. One case has already been discussed.

(*b*1) If $a = -3$, the system reduces to

$$\left(\begin{array}{cccc|c} 1 & 1 & 1 & -3 & A \\ 0 & -4 & 0 & 4 & C-A \\ 0 & 0 & -4 & 4 & B-A \\ 0 & 0 & 0 & 0 & D+C+A+B \end{array}\right).$$

In this case, a single compatibility condition is given by the last row

$$0 = A + B + C + D.$$

If this condition is not satisfied, the system is inconsistent (has no solution). If $A+B+C+D=0$, the system is

$$\left(\begin{array}{cccc|c} 1 & 1 & 1 & -3 & A \\ 0 & -4 & 0 & 4 & C-A \\ 0 & 0 & -4 & 4 & B-A \\ 0 & 0 & 0 & 0 & 0 \end{array}\right).$$

We may choose any value for u, say $u=c$, and infer from the third row that $-4z=-4c+B-A$. The second row now gives $-4y=-4c+C-A$. Finally the first row shows that

$$\begin{aligned} x &= -y-z+3u+A \\ &= \tfrac{1}{4}(-4c+C-A)+\tfrac{1}{4}(-4c+B-A)+3c+A \\ &= c+\tfrac{1}{2}A+\tfrac{1}{4}B+\tfrac{1}{4}C. \end{aligned}$$

This is an example of the *back-substitution* procedure. Since the value of the variable u can be chosen arbitrarily, we say that it is a *free variable*, and the solution list is

$$\begin{cases} x &= c+\frac{1}{2}A+\frac{1}{4}B+\frac{1}{4}C \\ y &= c+\frac{1}{4}A-\frac{1}{4}C \\ z &= c+\frac{1}{4}A-\frac{1}{4}B \\ u &= c. \end{cases}$$

($b2$) Finally, if $a \neq -3$ (and still $a \neq 1$), the system has a unique solution for each data A, B, C, and D. It is *regular.*

Let us observe *a posteriori* that the conditions found are quite natural. If $a=1$, the system is

$$\begin{cases} x+y+z+u &= A \\ x+y+z+u &= B \\ x+y+z+u &= C \\ x+y+z+u &= D, \end{cases}$$

whence the condition $A=B=C=D$. When $a=-3$ the system is

$$\begin{cases} x+y+z-3u &= A \\ x+y-3z+u &= B \\ x-3y+z+u &= C \\ -3x+y+z+u &= D, \end{cases}$$

and the sum of these equations is $0=A+B+C+D$. However, one cannot expect to *guess* the compatibility conditions for systems containing a large number of variables, hence the usefulness of the systematic elimination method.

1.2 Homogeneous Systems

1.2.1 A Chemical Reaction

Lord Rayleigh started his investigations on the composition of the atmosphere around 1894. He blew ammoniac (NH_3) and air on a red-hot copper wire and analyzed the result. Let us imitate him, and consider a typical reaction of the form

$$x\,NH_3 + y\,O_2 + z\,H_2 \rightarrow u\,H_2O + v\,N_2,$$

where the proportions $x, \ldots, v$ have to be found. (We have added hydrogen for mathematical interest, but we bet the reader to refrain from experimenting since such a mixture has an explosive character!) Equilibrium of N-atoms requires $x = 2v$. Similarly, equilibrium of hydrogen atoms requires $3x + 2z = 2u$ and finally, for oxygen, we get $2y = u$. As is required by the general method, we have first to adopt an order for the variables: Choose the order of occurrence in the chemical reaction, namely x, y, z, u, and v. Hence we write the system obtained in the form

$$\left\{\begin{array}{llllll} x & & & & -2v & = 0 \\ 3x & & +2z & -2u & & = 0 \\ & 2y & & -u & & = 0. \end{array}\right.$$

Now, observing that the right-hand sides are all zero, it is superfluous to include separators and the zeros that follow them, common to all equations: The first equation is abbreviated by the row $(1\ \ 0\ \ 0\ \ 0\ \ -2)$. The system of three equations is thus simply represented by the array

$$\begin{pmatrix} 1 & 0 & 0 & 0 & -2 \\ 3 & 0 & 2 & -2 & 0 \\ 0 & 2 & 0 & -1 & 0 \end{pmatrix}.$$

To eliminate x from the second equation on, subtract three times the first row from the second one. We obtain the equivalent system

$$\begin{pmatrix} 1 & 0 & 0 & 0 & -2 \\ 0 & 0 & 2 & -2 & 6 \\ 0 & 2 & 0 & -1 & 0 \end{pmatrix}.$$

Now exchange the second and third equations

$$\begin{pmatrix} 1 & 0 & 0 & 0 & -2 \\ 0 & 2 & 0 & -1 & 0 \\ 0 & 0 & 2 & -2 & 6 \end{pmatrix}.$$

This system is easily discussed since its second equation does not contain the first variable, while the third one does not contain the first two variables. The last equation is simply

$$2z - 2u + 6v = 0 \quad \text{or} \quad z - u + 3v = 0.$$

If we choose arbitrary values for u and v, say $u = a$ and $v = b$, we have to take

$$z = a - 3b.$$

The second equation then leads to $2y = a$, and the first one furnishes $x = 2b$. Thus, for each choice of a pair of values for u and v, there is one and only one *solution set*, or *list of solutions* given by

$$\begin{cases} x = & 2b \\ y = & \frac{1}{2}a \\ z = & a - 3b \\ u = & a \\ v = & b \end{cases} \quad \text{also denoted} \quad \begin{pmatrix} x \\ y \\ z \\ u \\ v \end{pmatrix} = \begin{pmatrix} 2b \\ \frac{1}{2}a \\ a - 3b \\ a \\ b \end{pmatrix}.$$

We consider such lists as mathematical objects, so that when we speak of *one* solution, we really mean a complete list: A *solution set.* In a similar vein, a linear system is a mathematical object, conveniently represented by the array of its coefficients. Entities considered by mathematicians are of different types, and if possible, a good notation should help to identify them.

Comment. The problem considered here is *homogeneous*, namely concerns proportions: If a solution is found, any *multiple* will also do. We can deal with numbers of atoms, or numbers of moles.[1] Two *basic solutions* appear. The first one corresponds to the choice $u = 2$, $v = 0$, hence $x = 0$ (no ammoniac); it corresponds to the elementary reaction

$$2\,H_2 + O_2 \to 2\,H_2O,$$

namely the synthesis of water. The other one—in which Lord Rayleigh was interested—corresponds to the choice $u = 6$, $v = 2$, hence $z = 0$ (no hydrogen, no danger in this case!) which corresponds now to the reaction

$$4\,NH_3 + 3\,O_2 \to 6\,H_2O + 2\,N_2.$$

Of course, one may *superpose* any multiples of these two basic reactions and obtain another possible one. This is reflected by the fact that the general solution of the system depends on two arbitrary *parameters* a and b: There are two *free variables* u and v.

1.2.2 Reduced Forms

In practice, systems containing hundreds or even thousands of equations and variables occur frequently: It is impossible to use tricks or guess work to solve

[1] Each mole contains approximately $0.60221367 \times 10^{24}$ atoms. At least, this is the currently accepted figure for the *Avogadro number*, namely the number of atoms in 12g. of the nucleid Carbon^{12}, or approximately the number of oxygen molecules O_2 in 32g. of this gas.

them. The alphabet is too small to code so many variables, so that we number them $x_1, x_2, x_3, \dots$ and thereby order them. Let us call n the number of variables, so that these unknown variables are labeled

$$x_1, x_2, x_3, \dots, x_n.$$

(Even if n is given explicitly, say $n = 1000$, there is an obvious advantage in the use of dots when we mention them!) The examples have shown the advantage of grouping the variables of equations in the left-hand side, the known quantities in the right-hand side, so we adopt the following definition.

Definition. *A* linear equation *in n variables is by definition a relation*

$$a_1x_1 + a_2x_2 + a_3x_3 + \cdots + a_nx_n = b,$$

where the literal coefficients *a_1, a_2, a_3,..., a_n, and b have some values. We abbreviate such an equation by the sequence of its coefficients, namely by the* row

$$(a_1 \ a_2 \ a_3 \ \dots \ a_n \mid b).$$

The separator *" | ", in place of the equality sign, distinguishes the left-hand from the right-hand sides of the equation. A* linear system *is a list consisting of a finite number of linear equations, each representing a condition to be satisfied by the unknown variables $x_1, \dots, x_n$.*

As we have seen in our second example, systems containing a number of equations different from the number of unknowns are important. A system containing a lot of equations in only two variables will usually have no solution. But a single equation in several variables has many solutions.

> ***The purpose of this chapter is to explain the* elimination procedure, *allowing to recognize when a linear system is* compatible, *and if so, determine its* solution(s).**

When a system is compatible, it is also important to be able to detect whether it has a *unique* or *many solutions.* Let us start by explaining this procedure when there are zeros after the separator " | ", namely when the right-hand sides of the linear equations are zero. Linear equations having a 0 after the separator are called *homogeneous.* One way of recognizing them is to substitute the value 0 for all variables and see if the equation is satisfied. Without reference to left-hand and right-hand sides, it is better to characterize homogeneity as follows.

Definition. *A linear system in n variables $x_1, \dots, x_n$ is* homogeneous *if it admits the* trivial solution $x_1 = 0$, $x_2 = 0$, ..., $x_n = 0$.

Since the linear homogeneous systems are compatible by definition, their study is simplified, and this is a good reason for discussing them first. The example of a chemical reaction treated in the preceding subsection has revealed an essential feature shared by all homogeneous systems:

➢ *Any multiple of a solution is again a solution*

➢ *The sum of two solutions is also a solution.*

The examples have also convinced us that a homogeneous system can always be transformed into an equivalent one (having the same solutions) where the nonzero coefficients form a *staircase pattern*. Ignoring the 0's in the right-hand sides, m homogeneous equations concerning n variables are described by a rectangular array of size m by n, and the discussion is easily made when the system has been brought into the following form

$$\left(\begin{array}{ccccccccc} p_1 & * & \cdots & & & \cdots & & & * \\ 0 & \cdots & p_2 & * & \cdots & & & & \\ & & 0 & \cdots & p_3 & * & & & \\ \vdots & & & & 0 & \cdots & & & \vdots \\ & & & & & \ddots & & & \\ & & & & & & p_r & * \;\cdots & * \\ 0 & \cdots & & & & & 0 & \cdots & 0 \\ \vdots & & & & & & \vdots & & \vdots \\ 0 & \cdots & & & & & 0 & \cdots & 0 \end{array}\right)\begin{array}{l} \left.\begin{array}{l} \\ \\ \\ \\ \\ \\ \end{array}\right\} r\text{: rank} \\ \left.\begin{array}{l} \\ \\ \\ \end{array}\right\} m-r \end{array}$$

where the coefficients $p_1, p_2, \ldots, p_r$ are nonzero: They are the *pivot values*, placed in *pivot positions*. By definition, the *rank* r is the number of nonzero lines: They are listed first. If $r < m$, the next $m - r$ lines are filled with zeros. The increasing integers

$$1 = j_1 < j_2 < \cdots < j_r,$$

are the indices of the *pivot columns*: They correspond to the *pivot variables* $x_{j_1}, x_{j_2}, \ldots, x_{j_r}$. By definition, the rank r is less than or equal to m and n

$$r \leqslant \min(m, n).$$

If $r < n$, there are $n - r$ nonpivot variables, called *free variables*. Starting from the last nonzero row, giving arbitrary values to the free variables, we can deduce the value of the last pivot variable x_r (thanks to $p_r \neq 0$). Working upwards, the given values of the free variables together with the previously found values for pivot variables, we can determine all pivot variables. This is the *back-substitution* procedure which leads to the general solution of the system. In particular, attributing the value 1 to one free variable, 0 to the others, we see that the linear homogeneous system has a nontrivial solution. This case certainly happens when $m < n$ (since $r \leqslant m$). It proves our first general result (it will play an important part in the next chapter).

Theorem. *A linear homogeneous system having more variables than equations admits a nontrivial solution.*[2] ■

[2]The character ■ stands for "end/absence of proof".

On the other hand, when $r = n$, there is no free variable, and the last row shows that $x_n = 0$. By back-substitution, we find successively $x_{n-1} = 0$, ... and finally $x_1 = 0$, so that the linear homogeneous system has only the trivial solution in this case.

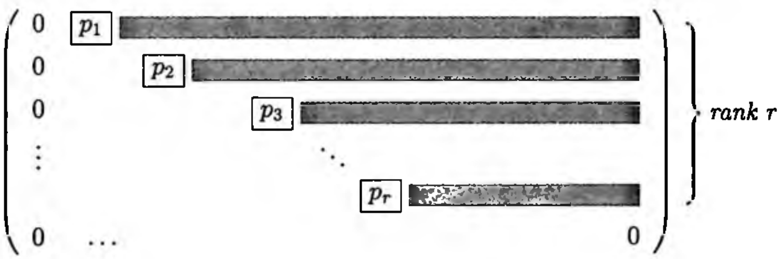
A *row-reduced array* is a special pattern where

- *The rows having only zeros come last,*
- *the first nonzero coefficients of rows come in increasing positions.*

Starting from *any* rectangular array, suitable transformations lead to such a form, where—in general—there might be some extra columns of zeros at the left. These first zero columns are absent when we start from a linear system, since there is no reason for introducing free variables that do not appear in the equations. But for the general discussion, we must also consider their possible presence: The first pivot column may not be the first column and we obtain an increasing sequence of pivot columns

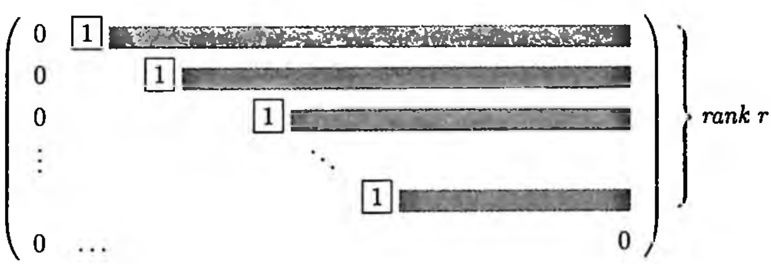
$$1 \leqslant j_1 < j_2 < \cdots < j_r \leqslant n.$$

Here is a picture of a row-reduced array (where for simplicity, we only include one first column and one last row of zeros)



ROW-REDUCED FORM

Here, the squares contain the pivot values $p_i \neq 0$ $(1 \leqslant i \leqslant r)$, while the grey rectangles may contain any entries. Multiplying the first row by $1/p_1$, the second by $1/p_2$, etc. we obtain an equivalent system where the pivot values are 1's.



ROW-ECHELON FORM

This is the *row-echelon form* (*echelon* refers to *unit pivot values*, but all reduced patterns have steps of unit height!). It is even possible to further simplify the system by requiring that all coefficients *above* a pivot position are 0's: Subtract a suitable multiple of the second row from the first one, suitable multiples of the third from the first and second ones, etc. (this does not destroy the main property of the reduced form, namely to have 0's in front of the pivots). This particular pattern is a *reduced* row-echelon form of the array A, that is conventionally abbreviated by rref(A).

$$\begin{pmatrix} 0 & \boxed{1} & 0 & 0 & 0 & \\ 0 & & \boxed{1} & 0 & 0 & \\ 0 & & & \boxed{1} & 0 & \\ \vdots & & & & \ddots & \\ & & & & \boxed{1} & \\ 0 & \dots & & & & 0 \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}0\\0\\0\\ \vdots\\ 1\end{matrix}}\right\} \text{\textit{rank} } r \\ \end{matrix}$$

REDUCED ROW-ECHELON FORM

This reduced row-echelon form corresponds to a system

$$\begin{cases} 0 + x_{j_1} \qquad\qquad\qquad + \quad \Sigma_1 = 0 \\ 0 \qquad + x_{j_2} \qquad\quad + \quad \Sigma_2 = 0 \\ 0 \qquad\qquad + x_{j_3} + \quad \Sigma_3 = 0 \\ \qquad\qquad\qquad\qquad \vdots \\ 0 \qquad\qquad\qquad + x_{j_r} + \Sigma_r = 0, \end{cases}$$

where $x_{j_1} = y_1$, $x_{j_2} = y_2, \ldots,$ and $x_{j_r} = y_r$ are the pivot variables, while the sums Σ_j only involve the free variables $y_{r+1}, \ldots, y_n$. (We use the capital Greek sigma Σ as an abbreviation for "sum"; since each row contains a possibly different one, we distinguish them by an index.) The solution of this system is obviously

$$\begin{cases} y_1 = x_{j_1} = -\Sigma_1 \\ y_2 = x_{j_2} = -\Sigma_2 \\ \quad \vdots \qquad \vdots \\ y_r = x_{j_r} = -\Sigma_r. \end{cases}$$

When $r < n$, the $n - r$ free variables can be given arbitrary values, and the system has infinitely many solutions. We say that the general solution depends on $n - r$ *parameters*.

Comment. Distinct sequences of operations may lead to row-reduced echelon forms. For example, if the first row starts by a 2, a possibility is to start by multiplying this row by $\frac{1}{2}$. If another row starts by a 1, another possibility is to exchange it with the first one to obtain a first pivot value 1. It is essential to realize that all methods end up with the same number of nonzero rows, so that *the rank r of a given rectangular array is well defined, independently of the method used to reach it*: We shall prove this *invariance* in Chapter 2. One can show that the indices of the pivots are well defined. Hence the distinction between *leading* variables and *free* variables is independent of the sequence of operations leading to a reduced form. But notice that this distinction depends on their order. For example, consider the homogeneous system in two variables $x + \xi = 0$, $x - \xi = 0$. If we adopt the order $x_1 = x$, $x_2 = \xi$, then $x_2 = \xi$ is a free variable; but if we reverse the order, ξ is the pivot variable.

1.3 Elimination Algorithm

An *algorithm*[3] is a systematic procedure leading to a solution of a certain class of problems. It is necessarily based on elementary operations which, taken individually may appear trivial but, furnish a nontrivial result when applied suitably and repeatedly. For example starting with two integers, the simple operation

subtract the small one from the large one

done repeatedly, leads to the greatest common divisor of these integers. More precisely, starting with a pair of distinct integers (m, n), proceed as follows:

$$\textit{replace } (m,n) \textit{ by } \begin{cases} (m-n, n) & \textit{if } m > n \\ (m, n-m) & \textit{if } m < n. \end{cases}$$

Continuing this procedure, we obtain a decreasing sequence of pairs of integers. After a finite number of steps, we shall reach a first pair (d, d) having two equal components: $d \geqslant 1$ is the greatest common divisor of m and n. This is the famous *Euclidean algorithm*. If, instead of integers, we start with a pair (a, b) of positive (real) numbers, the procedure may lead to a pair (d, d) after a finite number of steps: This is the case of commensurable numbers a and b. When the procedure never leads to such a pair (d, d), we say that a and b are incommensurable.

The *resolution algorithm* for linear systems starts as follows. Having ordered the variables, we group the monomials containing them in the left-hand side and replace each equation by the row of its coefficients. Thus the system is transformed into a rectangular array. When not all zero, the right-hand sides are listed after a separator, used as a reminder of the equality sign. Then elementary operations are performed in order to simplify the system: The goal is to reach a row-reduced form, from which the discussion (existence, uniqueness, and values of the variables) is easily carried out. With thousands of unknowns, all this would be done by a computer. But educated scientists should understand how and why it works. Let us explain it in more detail.

[3] From al-Khuwarizmi, Arab mathematician of the ninth century.

1.3.1 Elementary Row Operations

The elementary *row operations* used for transforming a linear system are:

1. ***Addition of a multiple of a row to another row***
2. ***Multiplication of a row by a nonzero number***
3. ***Exchange—or permutation—of two rows.***

Although the third type may be obtained using the first two types only (as we shall see), it is convenient to also treat it as an elementary operation.

The elementary row operations are *invertible*, hence they preserve the set of solutions of the system. Any sequence of row operations transforms the initial system into another one having the same solutions, called *equivalent system* for this reason. The goal is to reach a form in which the set of solutions is easily obtained. As we have seen with homogeneous systems, this is the case when the left part of the array—before the separator—is in *row-reduced* form:

- *The rows having only zeros before the separator come last*
- *The first nonzero coefficients (of rows) come in increasing order.*

The number of rows having a nonzero element before the separator is by definition the *rank* r of the system. The last $m - r$ rows, having only zeros before the separator, give *compatibility conditions* for the system. Indeed, when a row only has 0's in front of the separator, a nonzero after it leads to a contradiction. To solve the system, we proceed upwards, starting from the last nonzero row, attributing arbitrary values to the free variables—if any—deducing the corresponding value of the last pivot variable. The second last row now gives the value of the second last pivot variable (depending on the choices of values of the free variables, if any), and so on.

A general linear system containing m equations in n variables is said to have *size* $m \times n$ (read m by n). We use the following notation: An extra index is used to identify the equations It has the following form

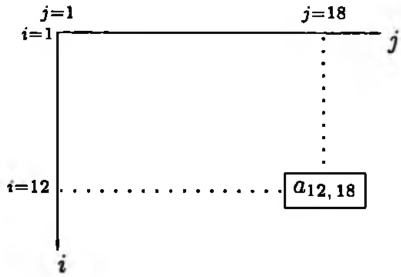
$$(S) \qquad \begin{cases} a_{11}x_1 & + \cdots + & a_{1n}x_n & = & b_1 \\ \vdots & & \vdots & & \vdots \\ a_{m1}x_1 & + \cdots + & a_{mn}x_n & = & b_m. \end{cases}$$

We also abbreviate it symbolically by $A\mathbf{x} = \mathbf{b}$ where the boldface font for "x" and "b" emphasizes that they represent *lists* instead of single numbers. At this point, this is only a symbolic representation for the list of equations, but it will soon appear to be a special case of *matrix multiplication*.

The system (S) is completely described by the *augmented array*

$$(A \mid \mathbf{b}) = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{array}\right).$$

Notice how the double indices are used: The first one indicates the row, the second one the column:



The size of the extended array is $m \times (n+1)$ (we always give the number of rows first), due to the presence of the list $\mathbf{b}$ in its last column. Suitable row operations on this array allow us to transform it into one having a first part (before the separators) in row-reduced or row-echelon form, say

$$A \sim U \text{ and } (A \mid \mathbf{b}) \sim (U \mid \mathbf{c}).$$

If U is in row-echelon form, here is how $(U \mid \mathbf{c})$ looks like.

$$\left(\begin{array}{ccccccccccc|c} 1 & * & \cdots & & & & \cdots & & & & * & c_1 \\ 0 & \cdots & & 1 & * & \cdots & & & & & * & c_2 \\ & & & 0 & \cdots & & 1 & * & \cdots & & * & c_3 \\ \vdots & & & & & & 0 & \cdots & & & & \vdots \\ & & & & & & & \ddots & & & & \vdots \\ & & & & & & & & 1 & * \cdots & * & c_r \\ & & & & & & & & 0 & \cdots & 0 & c_{r+1} \\ & & & & & & & & \vdots & & \vdots & \vdots \\ 0 & & \cdots & & & & & & 0 & \cdots & 0 & c_m \end{array}\right) \begin{array}{l} \left.\vphantom{\begin{array}{c}1\\1\\1\\1\\1\\1\end{array}}\right\} r \\ \left.\vphantom{\begin{array}{c}1\\1\\1\end{array}}\right\} m-r. \end{array}$$

The system has a solution precisely when

$$c_{r+1} = \cdots = c_m = 0 \quad \text{(compatibility conditions)},$$

or equivalently when

$$\operatorname{rank} U = \operatorname{rank}(U \mid \mathbf{c}).$$

When the system is compatible and $n > r$, it admits infinitely many solutions: The general solution depends on the arbitrary values chosen for the $n - r$ free variables. We say that it depends on $n-r$ *parameters.* When $r = m$, the system has *maximal* rank and it is always compatible. To summarize, we have

> ***uniqueness*** *when $r = n$: There is no free variable and there is* at most one *solution for each right-hand side* $\mathbf{b}$,

> ***existence*** *when $r = m$: There is no compatibility condition and a solution can be found for each right-hand side* $\mathbf{b}$,

> ***existence and uniqueness*** *(**regular system**) when $r = m = n$: The system has a unique solution for each right-hand side* $\mathbf{b}$.

Comments, Warnings

1. One cannot *simplify* by 0: from $1 \cdot 0 = 2 \cdot 0$ (true!), one cannot deduce $1 = 2$ (false!). Division by 0 produces an "ERROR 0" on a pocket calculator

Division by 0 is not a legal operation.

Multiplication by a number is always possible, but

Infinity is not a number.

Since a nonzero number a is invertible, it is legal to multiply by $a^{-1} = 1/a$, thus producing a division. Multiplication is a safe operation, division is not!

2. Solving an equation is not a matter of guessing. For example, to solve the (nonlinear) equation $x^2 = x$, we observe that it is equivalent to $x^2 - x = 0$, and to $x(x-1) = 0$. Here we see that $x = 0$ is a possibility. If $x \neq 0$, we may multiply by x^{-1} and obtain $x^{-1}x(x-1) = 0$, namely $x - 1 = 0$. Hence

$$x^2 = x \quad \text{implies} \quad x = 0 \text{ or } x = 1.$$

The following general ***Basic Principle*** ought to be remembered

$$ab = 0 \quad \textit{implies} \quad a = 0 \textit{ or } b = 0.$$

3. Several row operations may be performed in one step, and to save some writing, one often adds multiples of one row simultaneously to the other ones. But one has to keep in mind that row operations have to be *invertible*. For example, adding the second row to the first one, and simultaneously replacing the second row by the sum of the first two, is not a sequence of row operations (it obviously loses some information): Having added the second row to the first one, only this new first row may be used for further row operations (the old first row may be recovered by subtraction of the second from this new first row). A good practice consists in keeping a fixed underlined row, and add some of its multiples to other ones in order to simplify them.

1.3.2 Comparison of the Systems (S) and (HS)

A general linear system is represented by $A\mathbf{x} = \mathbf{b}$, or more explicitly

$$(S) \qquad \begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n = b_2 \\ \vdots \qquad\qquad \vdots \qquad \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n = b_m. \end{cases}$$

Using successive elementary row operations, we can transform it into $U\mathbf{x} = \mathbf{c}$ where U is row-reduced. The corresponding homogeneous system $A\mathbf{x} = \mathbf{0}$

$$(HS) \qquad \begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n = 0 \\ a_{21}x_1 + \cdots + a_{2n}x_n = 0 \\ \vdots \qquad\qquad \vdots \qquad \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n = 0 \end{cases}$$

is simultaneously equivalent to $U\mathbf{x} = \mathbf{0}$.

Let us examine the difference of two solutions $\mathbf{p} = (p_i)$ and $\mathbf{s} = (s_i)$ of (S). This difference $\mathbf{s} - \mathbf{p}$ is defined by $\mathbf{s} - \mathbf{p} = (s_i - p_i)$. It is obviously a solution $\mathbf{h} = (h_i)$ of (HS). Hence if we know a particular solution $\mathbf{p} = (p_i)$ of the linear system (S) (which is thus compatible), any other solution $\mathbf{s} = (s_i)$ has the form $\mathbf{s} = \mathbf{p} + \mathbf{h}$ where $\mathbf{h} = (h_i) = \mathbf{s} - \mathbf{p}$ is a solution of (HS). Hence $\mathbf{s} = (s_i)$ has the form

$$s_i = p_i + h_i \quad (1 \leqslant i \leqslant n) \textit{ where } (h_i) \textit{ is a solution of } (HS).$$

We have found the ***Fundamental Principle of Linear Algebra:***

> ***The general solution of a compatible linear system is the sum of any particular solution of (S) and the general solution of the associated homogeneous system (HS).***

To find a particular solution of (S), one may proceed by elimination, and select the solution corresponding to a zero value of all free variables. Let us recall the main property of the set of solutions of a homogeneous system:

- ➤ *If $\mathbf{s} = (s_i)$ is a solution, then $a\mathbf{s} = (as_i)$ is also one for any number a*
- ➤ *If $\mathbf{s} = (s_i)$ and $\mathbf{t} = (t_i)$ are solutions, then $\mathbf{s} + \mathbf{t} = (s_i + t_i)$ is also one.*

Variant. The theoretical discussion of the resolution of a linear system (S) can also be made by introduction of an extra variable z as follows. Let us consider the homogeneous system

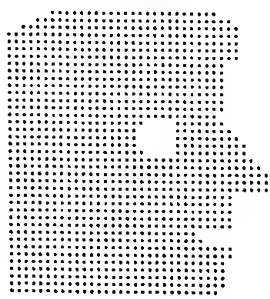
$$\begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n - b_1 z = 0 \\ a_{21}x_1 + \cdots + a_{2n}x_n - b_2 z = 0 \\ \vdots \qquad\qquad \vdots \qquad \vdots \quad \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n - b_m z = 0. \end{cases}$$

Then the solutions of the original system (S) correspond to the solutions of this homogeneous system having $z = 1$ (or $z \neq 0$, since a solution of (HS) may be multiplied by an arbitrary factor). When all solutions of this homogeneous system have $z = 0$, the original system is incompatible: (S) has no solution.

1.4 Appendix

1.4.1 Potentials on a Grid

Let us consider the following situation. In the plane $\mathbf{R}^2$, a certain bounded regular domain D is given (e.g. a disc, the interior of an ellipse, or a rectangle). We are looking for a potential inside D taking prescribed values on the boundary. To approach this physical problem, we introduce a square grid in the plane (having mesh of size $\varepsilon > 0$) and only keep the vertices of the squares having a nonempty intersection with the region D. We are left with a certain set of vertices P_i, which constitutes a discretization D_ε of D. Let us call *interior* vertices those having four neighbors (conveniently called North, East, South, and West) in D_ε, while the *boundary* vertices are those having less than four neighbors in D_ε. Here is an example of a discretization of a domain.



We are looking for a function f (potential) defined on the finite set D_ε, taking prescribed values on the boundary points and such that

> *$f(P)$ is the average of its four values at neighboring points for any interior point P.*

Let us number the points in an arbitrary way (starting from the interior ones), and introduce the variables $x_i = f(P_i)$ $(1 \leqslant i \leqslant N)$ for the unknown values of f at the corresponding interior points P_i. If the four neighbors of an interior point P_i are P_p, P_q, P_r, and P_s, there is a corresponding equation

$$x_i = \tfrac{1}{4}(x_p + x_q + x_r + x_s).$$

Here, $p = N(i)$ is the index of the northern neighbor of P_i, $q = S(i)$ is the index of the southern neighbor of P_i,... It may happen that all x_j are unknown, in which case we get a homogeneous equation

$$x_p + x_q + x_r + x_s - 4x_i = 0.$$

If, on the contrary, certain values are prescribed (because the corresponding points lie on the boundary), we get a nonhomogeneous equation. For instance, we may encounter an equation of the form

$$x_p + x_q + x_r - 4x_i = -b_s,$$

where b_s is the given value for the potential at a boundary point P_s ($s > N$). (Note that certain boundary values are irrelevant: Such are corner values, having no interior point as neighbor.) In any case, we can group the unknown variables in the left-hand side, while the known ones are gathered in the right-hand side. In this way, we obtain a linear system (S) for the variables x_i ($1 \leqslant i \leqslant N$). *We are going to show that this linear system is compatible, and has a unique solution for each data on the boundary.*

If there are N interior points P_i, the system contains N variables x_i and also N equations: We are going to show that (S) has maximal rank $r = N$. To prove this, we consider the associated homogeneous system (HS), simply obtained by requiring zero values on the boundary: In this case, it is enough to show that there is only one solution to the problem, namely the trivial one $x_i = 0$ for all indices i (corresponding to interior points P_i). Here is the crucial observation. For any solution set (x_i), select a variable x_j taking the maximal value (in a finite list, there is always a maximum). Since this value x_j is the average of the four values at its neighboring points, the only possibility is that these four values are equal, and equal to the maximal value. Iterating this observation at neighboring points, we eventually reach a boundary point where the value is 0. Hence the maximal value is itself 0. The same argument shows that the minimal value is 0. Finally, we see that all $x_i = 0$, which proves the claim. More generally, the mean value property shows that any solution takes values between its minimum and its maximum on the boundary. In other words, any solution reaches both a maximum and a minimum at a boundary point.

1.4.2 Another Illustration of the Fundamental Principle

Scenery: A river, a heap of peanuts, and a certain number of sleeping monkeys (in the shade of a palm tree!). Say there are N monkeys and x peanuts.

Action: A first monkey wakes up, counts the peanuts and finds that if he throws one into the river, which he does, the rest is divisible by N (isn't he smart!). He then eats his share and goes back to sleep (to the end of the story). Then a second monkey (as clever as the first one) wakes up —ignoring that another one has woken up before him—counts the peanuts and finds that if he throws a single one into the river—which he also does—the rest is divisible by N. He eats what he thinks is his share and goes back to sleep (also until the end of the story). And so on, until the Nth and last monkey, who makes the same observation, acts similarly.

Question. If N is given, find the smallest number of peanuts that is compatible with this story. For example, check that with 5 monkeys, an initial number of 3121 peanuts works. The successive remainders in this particular case are

$$2496\,,\ 1996\,,\ 1596\,,\ 1276\,,\ 1020.$$

ANSWER. Let x_i be the number of peanuts remaining when the first i monkeys have eaten what they thought was their share. We have $x_0 = x$ and then

$$x_1 = (x-1)\left(1-\tfrac{1}{N}\right), \quad \dots, \quad x_{i+1} = (x_i - 1)\left(1-\tfrac{1}{N}\right).$$

We find relations in the form

$$x_i = x_0\left(1-\tfrac{1}{N}\right)^i - A_i,$$

where A_i is independent from x. The resolution of the homogeneous system—A_i are all zero—is easy enough. Starting from an arbitrary x_0, one can compute successively x_1, x_2, ... The divisibility condition at the ith stage requires divisibility by N^i, and to end up in whole numbers, it is necessary to start with a multiple of N^N. Thus we write the general solution of the homogeneous system as

$$x_0 = cN^N\,,\ x_1 = \cdots$$

Integral values of c will lead to integral solutions of the homogeneous system, while other values of this parameter will lead to general solutions—not necessarily integral ones. There only remains to find a particular solution of the nonhomogeneous system. But I claim that

$$x = x_0 = 1 - N = x_1 = x_2 = \cdots = x_N$$

is one: Just play the game with negative numbers. Indeed, if there are $1-N$ peanuts in the heap (a debt), and we throw one away (thus increasing the debt by one), we end up with $-N$ peanuts. After eating his share (in this case, paying his part of the debt), the heap will again resume its size of $1-N$. And the next monkey does similarly. Now, the general solution of the nonhomogeneous system is the sum of this particular (negative) solution and of the general solution of the associated homogeneous system

$$x = 1 - N + cN^N.$$

The minimal positive one is obtained with $c = 1$

$$x_{\text{min}} = 1 - N + N^N.$$

For $N = 5$, we obtain $x_{\text{min}} = 1 - 5 + 5^5 = -4 + 5 \cdot 25^2 = 5 \cdot 625 - 4 = 3121.$

1.4.3 The Euler Theorem $f + v = e + 2$

The following experiment gives a plausible PROOF of the Euler theorem on the sphere.

Let the surface of a sphere be partitioned into f pools (faces), separated by e dams (edges). Suppose that each edge is common to two faces having among their vertices the two ends of this edge. In this proof, three *or more* faces may have a common vertex. We plan to irrigate the complete sphere by destruction of a minimal number of dams, starting with one single pool filled with water. At least one dam has to be broken to fill an empty pool. If we do this in the most economical way, exactly $f - 1$ dams have to be broken to completely flood the sphere. Having done that, we may count the number of intact ones. These will form a *tree*, namely a connected system of dams with no loop. But any such tree can be drawn in the following way:

- ➢ Start with the basic unit configuration containing 1 edge and 2 vertices
- ➢ Add successively branches, increasing simultaneously both the number of edges and the number of vertices by 1.

As we see, the iterative construction of any tree preserves the relation $e = v - 1$ at all steps. In particular, in our case we find

$$\begin{aligned} \text{number of broken edges} &= f - 1, \\ \text{number of intact edges} &= v - 1. \end{aligned}$$

Adding these relations, we find

$$e = \text{total number of edges } = f + v - 2.$$

This is the announced relation. ∎

Comment. Notice that on the surface of a sphere, any cycle of dams isolates a region: Whence the tree (or forest) structure of the intact dams after any complete flooding of the sphere. This is not the case on the surface of a torus where one equator does not separate two territories. In this case, a flooding of the complete surface my leave two cycles of dams intact. The corresponding Euler relation for any polygonal partition of a torus is $f + v = e$. Hence the linear system corresponding to a partition into pentagons and hexagons on a torus is homogeneous: It is the homogeneous system associated to the linear system obtained from the sphere. In this case, the number of pentagons is necessarily 0, while the number of hexagons is variable.

1.4.4 Fullerenes, Radiolarians

Fullerenes

Pure natural carbon can be found in several allotropic forms: Carbon powder, graphite, diamond, and as we now know, fullerenes of several types corresponding to stable molecules C_n in the form of tubes or spheres. The most famous one is the buckminsterfullerene C_{60}, which illustrates a decomposition of the sphere into hexagons and pentagons. It is by looking for linear molecules containing many carbon atoms in sidereal space that HAROLD W. KROTO (born 1939, professor at the University of Sussex, Brighton, G.-B.). finally understood the simple form that the carbon atoms can display in C_{60} (the actual discovery can be dated precisely 4.09.85: See NATURE, vol.318). Eventually, he found that these molecules are already produced—in small quantities—by pipe smokers! The 1996 Nobel prize in chemistry was indeed attributed to him and R. Curl, R.E. Smalley for their understanding of these beautiful molecules. In C_{60}, the carbon atoms are placed at the vertices of 12 pentagons, members of a partition of the sphere also containing 20 hexagons (think of a football ball!). The molecule C_{60} has diameter ≈ 10Å (1Å $= 10^{-10}$ m. represents roughly the diameter of an hydrogen atom). Hence the diameter of a molecule C_{60} is about 1 nanometer ($= 10^{-9}$ m.). It is now possible to synthesize rather inexpensively macroscopic quantities of the buckminsterfullerene C_{60} (purified at 99.5%). Other cage-like molecules containing only carbon atoms can be found or synthesized: C_{70} (played an important part at the beginning of the theory), $C_{240}, \ldots$ Long tubes of carbon atoms are promised a brilliant future! The term "fullerene" has been chosen by Kroto in honor of the American engineer and philosopher RICHARD BUCKMINSTER FULLER (1895-1983), who constructed geodesic domes, based on hexagonal and pentagonal decompositions of a hemisphere (US pavilion at the world exhibit in Montreal 1967, Union Tank building in Baton Rouge, Louisiana, etc.) In 2001, fullerenes were even found in rocks from the end of the Permian.

Radiolarian

This is a class of unicellular beings (protozoa belonging to marine plankton) having a skeleton in the shape of a polyhedral structure, allowing their pseudopodia to radiate through the pierced faces, most of them—not all—having a hexagonal shape, e.g. *Aulonia hexagona*. They are traditionally considered in the animal reign, since they can move and capture other small organisms (amoebaes, leukocytes). Their radiating thin feet allow them mobility (for capturing other microorganisms). The skeleton itself exhibits many hexagonal holes, a reason for the terminology *hexagonal radiolarian*. Nevertheless, each of them exhibits a few exceptional faces: Either they are perfect and only have twelve pentagonal holes, or they have extra heptagonal (rarely octagonal) ones.

1.5 Exercises

1. (*a*) Consider all possible repartitions of the surface of a sphere by curvilinear squares and triangles where each vertex is adjacent to *four* of these faces. Are there many possibilities? Is there a fixed number of triangles? Or squares? Does the cube lead to a special solution of the considered type? Is there a solution with squares only? (Repeat the discussion made for hexagons and pentagons in

this context.)
(*b*) Same as before for the repartitions of the surface of a sphere by curvilinear pentagons and triangles (where each vertex is still adjacent to four of these faces). Are there many possibilities? Is there a fixed number of triangles? Or pentagons? Is there a solution with triangles only?

2. (*a*) Consider all possible repartitions of the surface of a sphere by curvilinear squares and hexagons, where each vertex is adjacent to *three* of these faces. Are there many possibilities? Is there a fixed number of squares? or hexagons? Is there a solution with squares only?
(*b*) Same as before for the repartitions of the surface of a sphere by curvilinear triangles and octagons (where each vertex is still adjacent to three of these faces). Are there many possibilities? Is there a fixed number of triangles? or octagons? Is there a solution with triangles only?

3. Give a particular solution of the following linear system

$$(S) \qquad \begin{cases} y+z+w &= 2 \\ x+z+u &= 2 \\ x+y+u &= 2 \end{cases}$$

having $x = y = z$ and $u = w$. What is the general solution of (S)?

4. Consider the following linear system

$$\begin{cases} x_1 + 2x_2 - x_3 + 2x_4 &= a \\ x_1 - x_2 + x_3 - x_4 &= b \\ 4x_1 - x_2 + 2x_3 - x_4 &= 2. \end{cases}$$

For which values of a and b is it compatible? Find its general solution when it is compatible.

5. Let us consider functions f defined on the integers between 0 and a certain positive integer N, satisfying

$$f(n) = 1 + \text{average of } (f(n-1), f(n+1)) \quad (1 \leqslant n < N).$$

(a) Check that the function h defined by $h(n) = -n^2$ is a particular solution of the required functional equation.
(b) All functions g of the form $g(n) = An+B$ satisfy the associated homogeneous conditions.
(c) Deduce the solution f satisfying the two limit conditions $f(0) = 0$ and $f(N) = 0$.

6. Find correct coefficients $x, \dots, w$ for the chemical reaction

$$x\, Br^- + y\, H^+ + z\, MnO_4 \longrightarrow u\, BrO_3^- + v\, Mn^{2+} + w\, H_2O.$$

7. Same as before for the coupled reactions

$$\begin{aligned} *Fe + *OH^- &\longrightarrow *Fe(OH)_2 + *e^-, \\ *H_2O + *O_2 + *e^- &\longrightarrow *OH^-, \\ *Fe(OH)_2 + *O_2 + *H_2O &\longrightarrow *Fe(OH)_3. \end{aligned}$$

8. Consider the following linear system in n variables

$$\begin{cases} x_1 + x_2 &= a_1 \\ x_2 + x_3 &= a_2 \\ \vdots & \vdots \\ x_n + x_1 &= a_n. \end{cases}$$

Discuss completely the cases $n = 2$, 3, and 4. Can you generalize to any positive integer n?

9. Let M_1, $M_2, \ldots, M_n$ be n given points in the plane $\mathbf{R}^2$. When is it possible to find a closed polygonal line P_0, $P_1, \ldots$, P_{n-1}, $P_n = P_0$ such that M_i is the midpoint between P_{i-1} and P_i $(1 \leqslant i \leqslant n)$? When it is possible, are there many possibilities?

10. Let P_1, $P_2, \ldots, P_n$ be n given points in the space $\mathbf{R}^3$. Is it always possible to find disjoint balls B_i with center P_i $(1 \leqslant i \leqslant n)$ such that B_i is tangent to both B_{i-1} and B_{i+1}, where $B_0 = B_n$ and $B_{n+1} = B_1$. The problem is to find the radii of these balls, as a function of the distance of consecutive P_i's.

11. The equation of a plane in the usual space has the form

$$ax + by + cz = d,$$

where a, b, c, and d are parameters depending on the plane. Find all planes containing the points $P_1 = (1,1,1)$ and $P_2 = (1,2,3)$.

12. Are the following arrays

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

row-reduced? What is their rank?

13. How many free variables are there in the homogeneous system

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}?$$

14. What is the rank of the following homogeneous system in three variables

$$\begin{pmatrix} t^2 & t & 1 \\ t & 1 & t \\ 1 & t & t^2 \end{pmatrix}$$

as a function of the parameter t?

15. What is the rank of the following array

$$\begin{pmatrix} 1 & 2 & 3 & \dots & n \\ 2 & 3 & 4 & \dots & n+1 \\ \vdots & & & & \vdots \\ n & n+1 & n+2 & \dots & 2n-1 \end{pmatrix}?$$

16. Solve the following nonlinear system

$$\begin{cases} x^2yz &= 18 \\ xy^3z &= 24 \\ xyz^4 &= 6. \end{cases}$$

17. Is it possible to find α, β, and x such that

$$\begin{cases} \sin\alpha + \tan\beta - x^2 &= 2 \\ 2\sin\alpha + 2\tan\beta + x^2 &= 1 \\ -\sin\alpha - \tan\beta - x^2 &= 0? \end{cases}$$

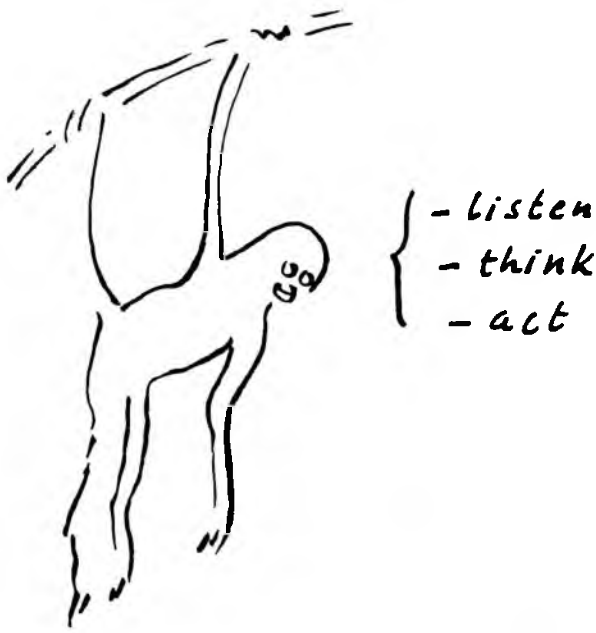
18. Find the simplest linear systems having many solutions, or no solution.

Notes

The elimination procedure is often called *Gaussian*, or *Gauss–Jordan elimination*. However, it was used since the antiquity by the Chinese, as reported by a manuscript of the third century of our era (see the book by Peter Gabriel listed in the references at the end of this volume). Hence it would be more accurate to call it the *fang-cheng algorithm*.

Keywords for Web Search

Aulonia hexagona (or hexagons)
Buckminster Fuller, fullerenes
Icosahedron, Platonic solids
www.mathworld.wolfram.com
Partial pivoting (row operations)
Fang–cheng algorithm (according to Chang Ts'ang)
Gaussian or Gauss–Jordan elimination



One row operation is particularly unsuitable in this situation: Which one?

Chapter 2

Vector Spaces

Linear equations in n variables are *mathematical objects*: They can be added and multiplied by numerical quantities. We say that they form a *vector space.*

2.1 The Language

2.1.1 Axiomatic Properties

Definition. Any set E consisting of mathematical objects which may be added, and multiplied by numbers, having the properties listed below is a vector space.

For simplicity, let us assume that the numbers involved are real numbers. However, not all the properties of the real number system $\mathbf{R}$ will be used, and one could equally well use only rational numbers (or complex numbers, to be introduced in Sec. 3.3.2). The properties implicitly used of numbers are listed at the end of this chapter. Here are the formal properties of addition and multiplication required in a vector space.

For any pair $\mathbf{x}, \mathbf{y}$ *in E, an element* $\mathbf{x}+\mathbf{y}$ *in E is well-defined, and this* sum *has the properties:*

1. $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z}$ (*associativity*).
2. $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$ (*commutativity*).
3. *There is a unique* $\mathbf{0}$ *in E such that* $\mathbf{x} + \mathbf{0} = \mathbf{x}$ *for all* $\mathbf{x}$ *in E.*
4. *Any* $\mathbf{x}$ *in E has a unique opposite* $\mathbf{x}'$ *in E such that* $\mathbf{x} + \mathbf{x}' = \mathbf{0}$.

On the other hand, for any pair consisting of a number a and an element $\mathbf{x}$ *in E, an element $a\mathbf{x}$ in E is well-defined, and this* multiplication *satisfies:*

5. $a(b\mathbf{x}) = (ab)\mathbf{x}$.
6. $a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y}$ (*distributivity*).

7. $(a+b)\mathbf{x} = a\mathbf{x} + b\mathbf{x}$ (*distributivity with respect to the sum of numbers*).

8. $1\,\mathbf{x} = \mathbf{x}$ (1 *denotes the unit number*).

For example, it is easy to see that $\mathbf{x}' = (-1)\mathbf{x}$ satisfies $\mathbf{x} + \mathbf{x}' = \mathbf{0}$. It is essential to observe that in general, the elements of a vector space cannot be multiplied together: In a vector space, the presence of an inner multiplicative law is not required (not forbidden either). The preceding list of axiomatic properties is not to be learnt by heart. Associativity, commutativity, distributivity,... are so natural that we shall hardly ever refer to them explicitly. In short, a vector space is a set in which finite sums of multiples of elements, called *linear combinations*, can be made.

The elements of a vector space are called *vectors* while numbers are called *scalars*. It is suitable to use different alphabets for them. For example, if ρ denotes the row $(2,3,-1)$, namely the homogeneous equation $2x_1+3x_2-x_3=0$, the multiple $a\rho$ denotes $(2a, 3a, -a)$, namely the equation $2ax_1+3ax_2-ax_3=0$. If $\rho' = (1,2,2)$ is another row of the same type (homogeneous equation in three variables), then

$$2\rho + \rho' = (4,6,-2) + (1,2,2) = (5,8,0)$$

represents the equation $5x_1 + 8x_2 + 0x_3 = 0$ of the same type. The zero row is $(0,0,0)$: It corresponds to the (trivial) equation in three variables

$$0x_1 + 0x_2 + 0x_3 = 0.$$

As a rule, the elements of an abstract vector space E will be typed with a boldface font as in $\mathbf{a}$, $\mathbf{x}$. This should help one to distinguish them from usual numbers $a, b, x, \dots$ In particular, one has to distinguish between the zero vector $\mathbf{0} \in E$ and the zero scalar 0. In specific examples, we may use a notation which is better adapted to the situation. For instance ρ may denote a row (of a specified length), $\vec{\mathbf{v}}$ a vector in the usual 3-space, f a function, etc. The notation should only be chosen in such a way as to suggest the correct interpretation.

2.1.2 An Important Principle

Here is a basic principle that follows from the axioms of vector spaces.

Proposition. *If a is a scalar and $\mathbf{x}$ is a vector, then*

$$a\mathbf{x} = \mathbf{0} \quad \textit{if and only if} \quad a = 0 \ \textit{or} \ \mathbf{x} = \mathbf{0}.$$

PROOF. For any $\mathbf{x} \in E$

$$0\mathbf{x} + 0\mathbf{x} = (0+0)\mathbf{x} = 0\mathbf{x},$$

and adding the opposite $(0\mathbf{x})'$ of $0\mathbf{x}$, we find $0\mathbf{x} = \mathbf{0}$. From this, we easily deduce $\mathbf{x}' = (-1)\mathbf{x}$ (simply written $-\mathbf{x}$): Indeed

$$\mathbf{x} + (-1)\mathbf{x} = (1 + (-1))\mathbf{x} = 0\mathbf{x} = \mathbf{0}.$$

For any scalar a, we have

$$a0 + a0 = a(0+0) = a0.$$

Adding the opposite of $a0$, we find $a0 = 0$. This already proves that if $a = 0$ or $\mathbf{x} = 0$, then $a\mathbf{x} = 0$. Conversely, if a is a nonzero scalar, it is invertible and

$$a^{-1}(a\mathbf{x}) = (a^{-1}a)\mathbf{x} = 1\mathbf{x} = \mathbf{x}.$$

Hence $a \neq 0$ and $a\mathbf{x} = 0$ implies $\mathbf{x} = a^{-1}(a\mathbf{x}) = a^{-1}0 = 0$. ■

2.1.3 Examples

(1) The set of linear homogeneous equations in n variables is a vector space. In fact, we have identified a linear homogeneous equation $a_1x_1 + \cdots + a_nx_n = 0$ to the row $(a_1, \ldots, a_n)$, and the set of rows of fixed length n is a vector space denoted by $\mathbf{R}_n$. Its elements are the rows

$$\rho = (a_1, \ldots, a_n), \quad \rho' = (a'_1, \ldots, a'_n), \quad \ldots$$

Linear combinations of rows are computed according to the rule

$$a\rho + \rho' = (aa_1 + a'_1, \ldots, aa_n + a'_n).$$

With rows $\rho_1, \ldots, \rho_m$, one may also consider the linear combinations

$$a_1\rho_1 + a_2\rho_2 + \cdots + a_m\rho_m,$$

where $a_1, a_2, \ldots, a_m$ are scalars. Such linear combinations can also be abbreviated in the form of a sum of the *generic term* $a_i\rho_i$ for the values of the index i between 1 and m:

$$a_1\rho_1 + a_2\rho_2 + \cdots + a_m\rho_m = \sum_{1\leqslant i\leqslant m} a_i\rho_i.$$

Another, closely related vector space, is the space consisting of the linear equations in n variables, represented by the rows $(a_1, \ldots, a_n \mid b)$. Since addition and multiplication of equations by scalars are computed termwise, this space is a splitting image of $\mathbf{R}_{n+1}$.

(2) A list of n scalars written vertically is called an *n-tuple*. For example, a solution set of an equation in n variables is an n-tuple. One should be careful about distinguishing rows of length n, and n-tuples which are columns of hight n. All n-tuples, with termwise addition and multiplication by numbers, form a vector space denoted by $\mathbf{R}^n$. The spaces $\mathbf{R}^n$ (for variable values of n) play a fundamental part in linear algebra. In a sense, $\mathbf{R}^n$ and $\mathbf{R}_n$ are symmetrical objects (see Chapter 9). The usefulness of the algebraic structure on n-tuples has already appeared when we presented the general solution of a linear system

in n variables as the *sum* of a particular solution of the linear system and the general solution of the associated homogeneous system:

$$\begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_n \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix} + \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{pmatrix} = \begin{pmatrix} p_1 + h_1 \\ p_2 + h_2 \\ \vdots \\ p_n + h_n \end{pmatrix}.$$

For $n = 1$, the space $\mathbf{R}^1$ can be identified with the particular vector space $\mathbf{R}$ consisting of the scalars.

(3) A linear homogeneous system containing m equations in n variables has been identified with an array of size $m \times n$. All arrays of fixed size $m \times n$ form a vector space if addition and multiplication by scalars are defined componentwise. The axiomatic properties required for a vector space are satisfied by these laws. Hence the rectangular arrays of fixed size $m \times n$ constitute a vector space $\mathbf{M}_{m,n}(\mathbf{R})$ (also denoted by $\mathbf{R}^m_n$). For example, the sum of two arrays of size 2×3 (2 rows and 3 columns) is defined by

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \end{pmatrix} = \begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} & a_{13}+b_{13} \\ a_{21}+b_{21} & a_{22}+b_{22} & a_{23}+b_{23} \end{pmatrix}.$$

The multiplication by a scalar is similarly defined by

$$a\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} = \begin{pmatrix} aa_{11} & aa_{12} & aa_{13} \\ aa_{21} & aa_{22} & aa_{23} \end{pmatrix}.$$

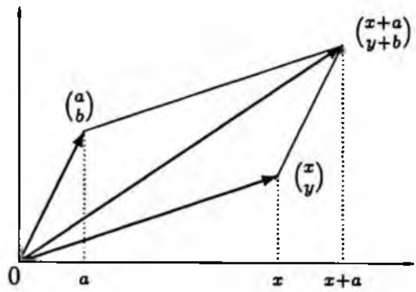
Of course, an array of size $1 \times n$ is a single row of length n, while an array of size $m \times 1$ is a single column of height m.

(4) A geometrical representation of the space $\mathbf{R}^2$ is given by a choice of *Cartesian coordinates* in the *Euclidean plane*. To the pair $\binom{x}{y}$ we associate the point P having coordinates x and y in this plane. In this way, pairs correspond 1-1 to points. The multiplication of a pair by a scalar, as the sum of two pairs, are defined componentwise:

$$a\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} av_1 \\ av_2 \end{pmatrix},$$

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} + \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} v_1 + w_1 \\ v_2 + w_2 \end{pmatrix}.$$

This makes it more intuitive to replace the point P by the vector $\vec{v} = \overrightarrow{OP}$. As a rule, elements of $\mathbf{R}^2$ will be represented by $\vec{v}$, $\vec{x}$, and so on. The addition now corresponds to the usual *parallelogram rule* for adding vectors.



A similar geometrical representation for the space $\mathbf{R}^3$ of triples of numbers occurs with a choice of *Cartesian coordinates* in the usual *Euclidean space*. Here, triples correspond 1-1 to points. Each triple $(x_i)_{i=1,2,3}$ corresponds to a point P having the coordinates x_i. We identify the vertically written triple (x_i) with the vector $\vec{\mathbf{v}} = \overrightarrow{OP}$, where P has for coordinates the three scalars x_i in the list. As a rule, the elements of this space will be represented by $\vec{\mathbf{v}}$, $\vec{\mathbf{x}}$, and so on.

If 4-tuples can be interpreted by vectors in space-time, it is difficult to build a representation for $\mathbf{R}^n$ for large values of n:

> *We simply consider that the elements of* $\mathbf{R}^n$ *are n-tuples, namely vertically written lists of n scalars.*

Comment. Observe that the usual Euclidean plane (or space) only becomes a vector space once an origin is chosen: This origin is then the zero $\mathbf{0}$ of this vector space. The addition (and multiplication by scalars) is not defined until an origin has been chosen.

2.1.4 Vector Subspaces

A subset V of a vector space E which is a vector space with the same laws as E (and the same zero vector as E), is a *subspace* of E.

Definition. *A* subspace *of a vector space E is a subset V that contains the zero vector $\mathbf{0} \in E$, and such that*

$$\mathbf{v} \textit{ and } \mathbf{w} \textit{ in } V \quad \textit{implies} \quad a\mathbf{v} + \mathbf{w} \textit{ in } V \textit{ for any scalar } a.$$

The zero vector of E alone, $V = \{\mathbf{0}\}$ is the smallest subspace of E; the whole space itself, $V = E$ is the largest. Hence, when $E \neq \{\mathbf{0}\}$, there are two trivial—extreme—examples of subspaces. We are mainly interested in *nontrivial* subspaces of a vector space E. But the trivial ones often occur as particular cases in general statements, and it would be awkward to exclude them *a priori*.

Examples. (1) Let us give the general form of the nontrivial subspaces of $\mathbf{R}^3$. First, we observe that the lines going through the origin (also called *homogeneous* lines) furnish infinitely many examples of subspaces of $\mathbf{R}^3$. Secondly, the planes containing the origin (*homogeneous* planes) furnish infinitely many other examples of subspaces of $\mathbf{R}^3$. As we shall see later, these are all the nontrivial subspaces of $\mathbf{R}^3$.

(2) Consider a system (HS) of homogeneous linear equations in n variables. The solutions of this system form a subset of the space $\mathbf{R}^n$ of n-tuples $(x_i)_{1\leqslant i\leqslant n}$. By definition, a linear system is homogeneous precisely when it admits the trivial solution (consisting of 0's only). In Sec. 1.2.2, we observed that the sum of two solutions, as well as the multiples of a solution, are again solutions of any homogeneous system. With the present terminology, we may say that the solutions of (HS) form a vector subspace of $\mathbf{R}^n$.

2.2 Finitely Generated Vector Spaces

2.2.1 Generators

Starting from a finite family of elements $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ of a vector space E, we can construct a vector subspace V of E as follows. Consider the subset consisting of all linear combinations of these vectors

$$\begin{aligned} V &= \text{set of linear combinations of } \mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n \\ &= \left\{\textstyle\sum_{1\leqslant i\leqslant n} x_i\mathbf{a}_i \ : \text{ any scalars } x_i\right\} \quad \text{subset of } E. \end{aligned}$$

This is a subspace since the trivial combination (all coefficients are 0) produces the zero vector of E, a multiple of a linear combination is again a linear combination:

$$a\,(x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n) = ax_1\mathbf{a}_1 + ax_2\mathbf{a}_2 + \cdots + ax_n\mathbf{a}_n$$

(by the axioms of vector spaces, valid in E), and similarly, the sum of two linear combinations

$$(x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n) + (y_1\mathbf{a}_1 + \cdots + y_n\mathbf{a}_n) = (x_1 + y_1)\mathbf{a}_1 + \cdots + (x_n + y_n)\mathbf{a}_n,$$

is again a linear combination. This subspace is called the *linear span* of the finite subset $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ (or of the family $(\mathbf{a}_i)_{1\leqslant i\leqslant n}$), and denoted by

$$V = \mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n) = \mathcal{L}(\mathbf{a}_i : 1 \leqslant i \leqslant n).$$

It is the smallest subspace containing $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$, since any subspace W of E, containing these elements, will also contain their linear combinations, hence contain $\mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n)$.

We also say that $V = \mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n)$ is *generated by* $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$, that these elements *generate*, or are *generators* of V.

Definition. *A vector space E is* finitely generated *when it has a finite generating family* $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$: $E = \mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n)$.

The commutativity of addition shows that

$$\mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) = \mathcal{L}(\mathbf{a}_2, \mathbf{a}_1, \ldots, \mathbf{a}_n),$$

and this subspace does not depend on the order in which the elements $\mathbf{a}_i$ are listed. It is also obvious that if $c \neq 0$

$$\mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) = \mathcal{L}(c\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n).$$

Hence this subspace $\mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)$ does not change if we replace any element $\mathbf{a}_i$ by one of its nonzero multiples $c\mathbf{a}_i$ $(c \neq 0)$.

Proposition. *With the previous notation, for any scalar c and $i \neq j$*

$$\mathcal{L}(\mathbf{a}_1, \ldots, \mathbf{a}_i, \ldots, \mathbf{a}_n) = \mathcal{L}(\mathbf{a}_1, \ldots, \mathbf{a}_i + c\mathbf{a}_j, \ldots, \mathbf{a}_n).$$

Before proving this proposition, let us recall the meaning of some set theoretic symbols (see Sec. 2.4.1).

$\mathbf{x} \in E$	means	$\mathbf{x}$ is an *element* of E,
		or $\mathbf{x}$ *belongs* to E,
$V \subset E$	means	V is a *subset* of E,
$E \supset V$	means	E *contains* V.

PROOF. Let us only prove the typical equality

$$\mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) = \mathcal{L}(\mathbf{a}_1 + c\mathbf{a}_2, \mathbf{a}_2, \ldots, \mathbf{a}_n).$$

By definition, for any scalar c we have

$$\mathbf{a}_1' = \mathbf{a}_1 + c\mathbf{a}_2, \mathbf{a}_2, \ldots, \mathbf{a}_n \in \mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n),$$

hence

$$\mathcal{L}(\mathbf{a}_1', \mathbf{a}_2, \ldots, \mathbf{a}_n) \subset \mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n).$$

Since conversely $\mathbf{a}_1 = \mathbf{a}_1' - c\mathbf{a}_2$, what we just observed also shows that

$$\mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) \subset \mathcal{L}(\mathbf{a}_1', \mathbf{a}_2, \ldots, \mathbf{a}_n),$$

so that finally, the equality

$$\mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) = \mathcal{L}(\mathbf{a}_1', \mathbf{a}_2, \ldots, \mathbf{a}_n)$$

follows by double inclusion. ∎

First Application. Let A denote a rectangular array of coefficients (corresponding to a linear system) containing m rows $\rho_1, \ldots, \rho_m$ of a certain type. In the vector space E of rows of this type, consider the *row space* of A, namely the subspace

$$\mathcal{L}(\text{rows of } A) = \mathcal{L}(\rho_1, \rho_2, \ldots, \rho_m)$$

generated by the rows of A. The preceding considerations show that a row operation does not change the row space:

Row operations preserve the row space of an array.

Performing successive row operations, we conclude that the row spaces of A and of any row-equivalent form are the same. In particular, if row operations are performed until a reduced form is obtained

$$A \sim A' \sim \cdots \sim U\text{: row-reduced form,}$$

the row space of A is the same as the row space of U:

$$\mathcal{L}(\text{rows of } A) = \mathcal{L}(\text{rows of } U).$$

Obviously, the row space of U is generated by its first nonzero rows.

Second Application. Let us interpret a linear system of size $m \times n$ in vector form, using the m-tuples

$$\mathbf{a}_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{mj} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} \in \mathbf{R}^m \quad (1 \leqslant j \leqslant n),$$

formed by the columns of its array of coefficients. We may rewrite the system in the equivalent form

$$x_1 \begin{pmatrix} a_{11} \\ \vdots \\ a_{m1} \end{pmatrix} + \cdots + x_n \begin{pmatrix} a_{1n} \\ \vdots \\ a_{mn} \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix},$$

or more simply

$$(S) \qquad x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = \mathbf{b}.$$

(1) The system (S) is *compatible* precisely when $x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = \mathbf{b}$ holds for suitable values of the coefficients x_i, namely when $\mathbf{b}$ *is* a linear combination of the vectors $\mathbf{a}_j$:

$$\mathbf{b} \in \mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n).$$

Since the inclusion $\mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) \subset \mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n, \mathbf{b})$ is obvious, (S) is compatible precisely when

$$\mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) = \mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n, \mathbf{b}).$$

(2) The same system (S) can be solved *for all data* $\mathbf{b} \in \mathbf{R}^m$ precisely when the linear span of $\mathbf{a}_1, \ldots, \mathbf{a}_n$ is the whole space

$$\mathcal{L}(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) = \mathbf{R}^m,$$

namely when the m-tuples $\mathbf{a}_1, \ldots, \mathbf{a}_n$ form a set of generators of $\mathbf{R}^m$.

(3) Assume now that (S) has infinitely many solutions (for a suitably given right-hand side $\mathbf{b}$), but all have the same value x_j for some fixed j (recall Example 1.1.2, where the number of pentagons $x = 12$ had a fixed value). To fix ideas, let us assume that x_1 has the same value in all solutions of (S). By the basic principle of linear algebra, all solutions are obtained from a particular one, by addition of a solution of

$$(HS) \qquad x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = \mathbf{0}.$$

Hence x_1 has a fixed value in all solutions of (S) when all solutions of (HS) have a zero value for x_1. This simply means that $\mathbf{a}_1$ is not a linear combination of $\mathbf{a}_2, \ldots, \mathbf{a}_n$. Indeed, any expression

$$\mathbf{a}_1 = c_2\mathbf{a}_2 + \cdots + c_n\mathbf{a}_n,$$

corresponds to a relation

$$\mathbf{a}_1 - c_2\mathbf{a}_2 - \cdots - c_n\mathbf{a}_n = \mathbf{0},$$

hence to a solution of (HS) having a first coefficient $x_1 = 1 \neq 0$. In other words, x_1 has a fixed value in all solutions of (S) precisely when $\mathbf{a}_1$ is not a linear combination of $\mathbf{a}_2, \ldots, \mathbf{a}_n$, namely

$$\mathbf{a}_1 \notin \mathcal{L}(\mathbf{a}_2, \ldots, \mathbf{a}_n).$$

More generally, x_j has a fixed value in all solutions of (S) precisely when $\mathbf{a}_j$ is not a linear combination of the other m-tuples $\mathbf{a}_i$:

$$\mathbf{a}_j \notin \mathcal{L}(\mathbf{a}_1, \ldots, \mathbf{a}_{j-1}, \mathbf{a}_{j+1}, \ldots, \mathbf{a}_n).$$

2.2.2 Linear Independence

By the basic principle of linear algebra, a linear system

$$(S) \qquad x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = \mathbf{b},$$

has at most one solution when the associated homogeneous system

$$(HS) \qquad x_1\mathbf{a}_1 + \cdots + x_n\mathbf{a}_n = \mathbf{0},$$

only has the trivial solution $x_1 = 0, \ldots, x_n = 0$. This happens when no $\mathbf{a}_j$ can be written as a linear combination of the other $\mathbf{a}_i$'s. Indeed, if $\mathbf{a}_j$ is a linear combination of the other $\mathbf{a}_i$'s, say

$$\mathbf{a}_j = c_1\mathbf{a}_1 + \cdots + c_{j-1}\mathbf{a}_{j-1} + c_{j+1}\mathbf{a}_{j+1} + \cdots + c_n\mathbf{a}_n,$$

then (HS) has a solution set with $x_j = 1$, hence a nontrivial solution. It is natural to say that the m-tuples $\mathbf{a}_1, \dots, \mathbf{a}_n$ are *linearly independent* when none is a linear combination of the others. As we have just observed, a linear system (S) has at most one solution when its columns, the m-tuples $\mathbf{a}_i$ $(1 \leqslant i \leqslant n)$, are linearly independent. This suggests a general definition, valid in any vector space.

Definition. *A finite family $(\mathbf{v}_i)_{1 \leqslant i \leqslant n}$ of elements of a vector space E is* linearly independent *when only the trivial linear combination of these elements gives the zero vector, namely when:*

$$\textstyle\sum_{1 \leqslant i \leqslant n} x_i \mathbf{v}_i = x_1 \mathbf{v}_1 + \cdots + x_n \mathbf{v}_n = \mathbf{0} \quad \textit{implies} \quad x_i = 0 \textit{ for all } i.$$

An infinite family of vectors is independent *when all its finite subfamilies are independent. When this is not the case, the family is* linearly dependent: *There is a nontrivial* dependence relation $x_1 \mathbf{v}_1 + \cdots + x_n \mathbf{v}_n = \mathbf{0}$, *having at least one nonzero coefficient.*

If the coefficient x_j in $x_1 \mathbf{v}_1 + \cdots + x_n \mathbf{v}_n = \mathbf{0}$ is nonzero, we can solve

$$\mathbf{v}_j = \textstyle\sum_{i \neq j} (-x_i / x_j) \mathbf{v}_i$$

and the corresponding vector $\mathbf{v}_j$ is a linear combination of the other $\mathbf{v}_i$'s. If a finite sum $\sum_i x_i \mathbf{v}_i = \mathbf{0}$ has a nonzero coefficient, then one vector is a linear combination of the others: But we do not know *a priori* which one. One technical advantage of the preceding definition is its symmetry: To prove that the n vectors $\mathbf{v}_i$ are independent, instead of having to show that

$$\begin{array}{lll} \mathbf{v}_1 & \text{is not a linear combination of} & \mathbf{v}_2, \mathbf{v}_3, \dots \\ \mathbf{v}_2 & \text{is not a linear combination of} & \mathbf{v}_1, \mathbf{v}_3, \dots \\ & \vdots & \\ \mathbf{v}_n & \text{is not a linear combination of} & \mathbf{v}_1, \dots, \mathbf{v}_{n-1}, \end{array}$$

we simply have to show that

$$x_1 \mathbf{v}_1 + \cdots + x_n \mathbf{v}_n = \mathbf{0} \quad \text{implies} \quad x_1 = 0, \dots, x_n = 0.$$

Comment. If a subset of the vector space E is linearly independent, it does not contain the vector $\mathbf{0} \in E$. Indeed, $1\,\mathbf{0} = \mathbf{0}$ is a nontrivial dependence relation. This leads to the following observation: In the trivial vector space $E = \{\mathbf{0}\}$, the only linearly independent subset is the empty set (an exceptional case!). From now on, we shall simply write 0 instead of $\mathbf{0}$ for the zero vector, relying on the reader for the proper interpretation.

Example. The nonzero rows of a row-reduced array are independent. If

$$U = \begin{array}{cc} \begin{pmatrix} p_1 & * & & & & \\ 0 & \cdots & p_2 & * & & \\ \vdots & & 0 & \cdots & p_r & * \\ \vdots & & \vdots & & 0 & \cdots \end{pmatrix} & \begin{array}{c} \rho_1 \\ \rho_2 \\ \rho_r \\ \vdots \end{array} \end{array}$$

with nonzero pivots p_i, a linear combination $\sum a_i \rho_i$ vanishes only if $a_1 = 0$: The first coefficient of this row is indeed $a_1 p_1$. Taking into account this fact, one can next look at the coefficient of index given by the second pivot, and see that $a_2 = 0$. Continuing in this way, we prove that all coefficients are zero.

2.2.3 The Dimension

Having discussed the notions of generating subset and of linearly independent subset, we combine the two. It is obvious that if we add elements to a generating subset of a space E, the enlarged set will *a fortiori* generate E. On the other hand, removing elements from a linearly independent set preserves linear independence.

Theorem. Assume that $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ *are* n *linear combinations of certain elements* $\mathbf{w}_1, \ldots, \mathbf{w}_m$ *of a vector space, namely*

$$\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \in \mathcal{L}(\mathbf{w}_1, \ldots, \mathbf{w}_m).$$

If $n > m$, *then the set* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ *is linearly dependent.*

PROOF. Let us start with n vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$, in the subspace generated by $\mathbf{w}_1, \ldots, \mathbf{w}_m$, where $n > m$. Let us write these linear combinations $\mathbf{v}_j$ more explicitly as

$$\mathbf{v}_j = a_{1j}\mathbf{w}_1 + a_{2j}\mathbf{w}_2 + \cdots + a_{mj}\mathbf{w}_m \quad (1 \leqslant j \leqslant n).$$

Now, let us form linear combinations of these

$$\begin{array}{rcl}
x_1\mathbf{v}_1 & = & x_1 a_{11}\mathbf{w}_1 + x_1 a_{21}\mathbf{w}_2 + \cdots + x_1 a_{m1}\mathbf{w}_m \\
 & \vdots & \\
x_n\mathbf{v}_n & = & x_n a_{1n}\mathbf{w}_1 + x_n a_{2n}\mathbf{w}_2 + \cdots + x_n a_{mn}\mathbf{w}_m \\
\hline
\sum_j x_j\mathbf{v}_j & = & (x_1 a_{11} + \cdots + x_n a_{1n})\mathbf{w}_1 \\
 & & \qquad + \quad \cdots \quad + \\
 & & \qquad (x_1 a_{m1} + \cdots + x_n a_{mn})\mathbf{w}_m.
\end{array}$$

To obtain zero with such a linear combination, we can simply choose the coefficients x_i solution of the homogeneous linear system

$$\begin{cases}
x_1 a_{11} + x_2 a_{12} + \cdots + x_n a_{1n} & = 0, \\
 & \vdots \\
x_1 a_{m1} + x_2 a_{m2} + \cdots + x_n a_{mn} & = 0.
\end{cases}$$

Since this system has more variables than equations ($n > m$), it has a nontrivial solution: The vectors $\mathbf{v}_j$ $(1 \leqslant j \leqslant n)$ are linearly dependent. ■

The preceding basic result may be reformulated as:

> *Any family having more elements than a generating set is linearly dependent,*

or equivalently:

> *Any* linearly independent *subset of* $\mathcal{L}(\mathbf{w}_1, \ldots, \mathbf{w}_m)$ *has at most* m *elements.*

Definition. *A* basis *of a vector space is a* linearly independent *and* generating *family.*

Let us consider *finitely generated* vector spaces, namely vector spaces E in which there is a finite family $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of elements such that $E = \mathcal{L}(\mathbf{v}_1, \ldots, \mathbf{v}_n)$. If this family is linearly dependent, one of its elements can be expressed as a function of the others, and deleting it, we obtain a set of generators having one less element. Continuing in this way, we finally reach a basis of E. We have thus proved the first part of the following basic theorem.

Theorem. *Let E be a finitely generated vector space. Then E has a finite basis. Two bases of E have the same number of elements.*

PROOF. Take two bases A and B of E, and let $\operatorname{card} A$, $\operatorname{card} B$ denote their respective number of elements. Then

$$\begin{aligned} A \text{ generates and } B \text{ independent} \quad &\Longrightarrow \quad \operatorname{card} B \leqslant \operatorname{card} A, \\ B \text{ generates and } A \text{ independent} \quad &\Longrightarrow \quad \operatorname{card} A \leqslant \operatorname{card} B. \end{aligned}$$

This proves $\operatorname{card} A = \operatorname{card} B$. ∎

Definition. *The common number of elements in all bases of a finitely generated vector space E is called* dimension *of E, and is denoted by* $\dim E$.

To be remembered! If the vector space E has dimension n, then any independent subset of E containing n elements, or any generating set containing n elements, is a basis of E. The existence of a basis in a finitely generated vector space E can be obtained in two complementary ways. First, a basis of E is a *minimal generating set* of this space. From any generating set, it is possible to extract a basis of E. Second, a basis is also a *maximal linearly independent subset* of E. Starting from any independent set in E, we may build a basis by successive introduction of elements which are not linear combinations of the previous ones. In a finitely generated vector space, this procedure will eventually furnish a basis, since any independent set has at most as many elements as there is in a generating subset.

Examples. (1) *The vector space $E = \mathbf{R}^n$ has dimension n.* To prove this, we have to find n elements forming a basis of this space. I claim that the following

n-tuples

$$\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \quad \dots, \quad \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

constitute a basis of the space. Let us make arbitrary linear combinations of these vectors. By definition

$$x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n = \begin{pmatrix} x_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ x_2 \\ \vdots \\ 0 \end{pmatrix} + \cdots + \begin{pmatrix} 0 \\ 0 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

Hence taking suitable coefficients x_i, we can obtain any n-tuple: These elements $\mathbf{e}_i$ make up a set of generators of $\mathbf{R}^n$. Moreover, a linear combination of these can furnish the zero n-tuple only if all coefficients x_i vanish: They are independent. In particular for $n = 1$, we see that the field of scalars is a one-dimensional vector space: Any nonzero element is a basis of this space. In a vector space $E \neq \{0\}$, there are infinitely many bases. However, the precedent basis of $\mathbf{R}^n$ is the most natural one, and is therefore called *the canonical basis* of $\mathbf{R}^n$. In an abstract vector space, there is usually no way of selecting a natural basis among all possible ones.

(2) In a similar vein, consider the vector space $\mathbf{R}^m_n$ consisting of arrays of size $m \times n$. It has a canonical basis $(E_{ij})_{1\leqslant i\leqslant m,\, 1\leqslant j\leqslant n}$, consisting of the arrays

$$E_{11} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & & \\ 0 & 0 & 0 & & \\ \vdots & & & \ddots & \vdots \\ 0 & & & \dots & 0 \end{pmatrix}, \quad E_{12} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & & \\ 0 & 0 & 0 & & \\ \vdots & & & \ddots & \vdots \\ 0 & & & \dots & 0 \end{pmatrix}, \quad \dots$$

where E_{ij} has only one nonzero coefficient—this being a 1—placed at the intersection of the ith row and jth column. Hence this vector space has dimension mn.

Fundamental Application. Let A be any rectangular array of coefficients. Using row operations, we can find a reduced row-equivalent form of A, say $A \sim U$. We have seen $\mathcal{L}(\text{rows of } A) = \mathcal{L}(\text{rows of } U)$. Now, the nonzero rows of U form a system of generators of this space. Since they are independent, they constitute a *basis* of the row-space:

$$r = \dim \mathcal{L}(\text{rows of } A).$$

Two procedures $A \sim U$ and $A \sim U'$ leading to row-reduced forms of A furnish two bases of $\mathcal{L}(\text{rows of } A)$, hence have the same number of elements: $r = r'$.

This proves that *the rank r is independent of the particular method of reduction used to find it*, and *the row-rank of A is unambiguously defined by*

$$\operatorname{rank} A = \dim \mathcal{L}(\text{rows of } A).$$

This is a first achievement of the language of linear algebra. Let us recall its filiation:

➢ *If a homogeneous linear system has more variables than equations, then it has a nontrivial solution*

➢ *Any independent set in $\mathcal{L}(\mathbf{w}_1, \ldots, \mathbf{w}_m)$ has at most m elements*

➢ *Two bases of a finitely generated vector space have the same number of elements.*

Let us come back to the subspace $V \subset \mathbf{R}^n$ consisting of the solutions of the homogeneous system $A\mathbf{x} = 0$: *It has dimension $n - r$.* Indeed, attributing successively the values $1, 0, \ldots, 0$, and then $0, 1, 0, \ldots, 0$, etc. to the free variables, we find a basis of V. This proves that the sum

$$\operatorname{rank} A + \dim V = r + (n - r) = n$$

is the same for all arrays A of the same size $m \times n$. This is a first form of the *rank-nullity theorem*, to be proved in Sec. 4.3.1.

***Theorem* (*Incomplete Basis*).** *Let $E = \mathcal{L}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m)$ be a finitely generated vector space, and S a linearly independent subset in E. Then there is a basis of E consisting of S and certain $\mathbf{v}_j$'s.*

PROOF. Start from an independent subset $S \subset E$: If it does not generate E, then at least one among the $\mathbf{v}_j$'s is not in the subspace $\mathcal{L}(S)$, and we consider the independent set $S' = S \cup \{\mathbf{v}_j\}$. After at most m such adjunctions, we obtain a maximal independent set, hence a basis of E of the required form. ■

2.3 Infinite-Dimensional Vector Spaces

The language of linear algebra is not restricted to finitely generated spaces.

For example, the set $E = \mathcal{F}(\mathbf{R}; \mathbf{R})$ of all functions $\mathbf{R} \to \mathbf{R}$ is a vector space. The zero function is the function that vanishes identically

$$f = 0 \quad \Longleftrightarrow \quad f(x) = 0 \textit{ for all } x \in \mathbf{R}.$$

We add real-valued functions on $\mathbf{R}$ as follows

$$(f + g)(x) = f(x) + g(x),$$

and multiply them by scalars similarly

$$(af)(x) = af(x).$$

(It turns out that we can also multiply two functions, hence speak of a multiplication law in this vector space; but this possibility is irrelevant here.) Hence a function $f : \mathbf{R} \to \mathbf{R}$ may also be called *vector*, when considered as an element of the vector space $\mathcal{F}(\mathbf{R}; \mathbf{R})$. This vector space is huge, and it is more reasonable to work with subspaces. For example, we learn in a first calculus course that the zero function is continuous, the sum of two continuous functions is continuous, and so are the multiples of continuous functions. Hence we may say that the subset of *continuous* functions $C(\mathbf{R}; \mathbf{R})$ is a subspace of $\mathcal{F}(\mathbf{R}; \mathbf{R})$.

For any subset A of a vector space E, the linear span $\mathcal{L}(A)$ of A is the smallest subspace of E containing A. It consists of the linear combinations—finite sums of multiples—of elements of A. If $(\mathbf{a}_i)_{i \in I}$ is any family in A, the notation

$$\sum_{\text{finite}} x_i \mathbf{a}_i$$

represents a finite sum, obtained by taking *at most finitely many* $x_i \neq 0$. These are the *linear combinations* that are considered in linear algebra: *Remember*

Infinite sums of nonzero elements are not defined in vector spaces.

The following properties of a vector space E are equivalent by definition:

(*i*) *E is not finitely generated*

(*ii*) *For each finite subset $A \subset E$, there is an $\mathbf{x} \in E$, $\mathbf{x} \notin \mathcal{L}(A)$*

(*iii*) *For each positive integer n, E contains independent subsets of cardinality greater than n.*

When they are satisfied, we say that E is *infinite dimensional*.

2.3.1 The Space of Polynomials

The simplest, and probably also most natural infinite-dimensional space is the space of polynomials.

Definition. *A* polynomial *in one variable x is a finite sum of multiples of the monomials x^n $(n \in \mathbf{N})$.*

By definition, a polynomial is a linear combination $f = \sum_{\text{finite}} a_i x^i$ of the integral powers of the variable x

$$x^0 = 1,\ x,\ x^2,\ \dots,\ x^n,\ \dots$$

For any nonzero polynomial f there is a maximal power in $\sum_{\text{finite}} a_i x^i$ that occurs with nonzero coefficient: This maximal power is the *degree* of f. Hence

$$\deg\left(\sum\nolimits_{\text{finite}} a_i x^i\right) = n \quad \Longleftrightarrow \quad a_n \neq 0 \text{ and } a_i = 0 \text{ for all } i > n.$$

Here are some polynomials

$$x^3 - 10x, \quad x^n + 1, \quad \frac{x^n - 1}{x - 1} = 1 + x + \cdots + x^{n-1}.$$

The degree of the first is 3, the degree of the second is n, and the degree of the last is $n-1$. But the following expressions are *not* polynomials

$$x^2 + x^{-2} = x^2 + 1/x^2, \quad x + x^{\frac{1}{2}}, \quad 1 + x + \cdots + x^n + \cdots$$

We learn in a calculus course that the last infinite series may be identified with $1/(1-x)$ when $|x| < 1$: It is not a polynomial.

Each polynomial $f = \sum_{\text{finite}} a_i x^i$ defines a *polynomial function*

$$x \longmapsto f(x) = \sum\nolimits_{\text{finite}} a_i x^i \ : \quad \mathbf{R} \longrightarrow \mathbf{R}$$

with which we may identify it,[1] since if $\sum_{\text{finite}} a_i x^i = \sum_{\text{finite}} b_i x^i$ for infinitely many values of x, then $a_i = b_i$ for all i.

The addition of linear combinations corresponds to the addition of functions (namely, adding their values), and a multiplication of a linear combination by a scalar corresponds to the multiplication of the corresponding function by the same scalar. (It turns out that we can also multiply together two polynomial functions, but this possibility is irrelevant here.) The definition of polynomials makes it obvious that a *basis* of the space of polynomials is furnished by the basic family

$$x^0 = 1,\ x,\ x^2,\ \ldots\ ,\ x^n,\ \ldots$$

A finite linear combination of the powers x^i vanishes by definition only when all coefficients vanish, hence this family is independent. On the other hand, the space of polynomials is generated by these powers

$$\Pi = \mathcal{L}(x^i \ : \ i \in \mathbf{N}).$$

The powers $i < n$ generate an nth-dimensional subspace

$$\Pi_{<n} = \mathcal{L}(1, x, \ldots, x^{n-1}) = \mathcal{L}(x^i : 0 \leqslant i < n).$$

We shall adopt the convention that the degree of the zero polynomial is less than 0. The inequality

$$\deg(f+g) \leqslant \max(\deg f, \deg g),$$

is then true in all cases (even when $g = -f$), and $\{f \in \Pi : \deg f < n\}$ contains the zero polynomial: We have

$$\Pi_{<n} = \{f \in \Pi : \deg f < n\}.$$

This sequence of subspaces of the space of polynomials starts by

$$\begin{aligned} \Pi_{<0} = \{0\} \quad &: \quad \text{The trivial subspace consisting of 0 only,} \\ \Pi_{<1} = \mathbf{R} \quad &: \quad \text{The subspace of constant polynomials,} \\ \Pi_{<2} \quad &: \quad \text{The subspace of polynomials } a + bx. \end{aligned}$$

[1] Here we use the fact that the scalar field is *infinite*. Recall that a polynomial f of degree $n \geqslant 1$ has at most n *roots*: There are at most n values of x such that $f(x) = 0$.

It increases indefinitely

$$\{0\} \subset \mathbf{R} \subset \Pi_{<2} \subset \Pi_{<3} \subset \cdots \subset \Pi_{<n} \subset \cdots \subset \Pi.$$

The expressions $a + bx + cx^2$ for some scalars a, b, and c make up the subspace $\Pi_{<3}$, of dimension 3. The *quadratic polynomials*, are the polynomials of degree 2, namely the expressions $a + bx + cx^2$ where $c \neq 0$. Note that the subset of quadratic polynomials *is not* a subspace of Π since the sum of two polynomials of degree 2 may have degree less than 2: For example

$$(x^2 + x + 1) + (-x^2 + 1) = x + 2$$

has degree 1. (The quadratic polynomials constitute the set theoretic complement of $\Pi_{<2}$ in $\Pi_{<3}$).

Remark. The powers $i \geqslant n$ generate an infinite-dimensional subspace

$$W_n = \mathcal{L}(x^n, x^{n+1}, \ldots) = \mathcal{L}(x^i : i \geqslant n).$$

The *order* of a nonzero polynomial $\sum_{\text{finite}} a_i x^i$ is by definition the smallest power i for which the coefficient $a_i \neq 0$. Hence

$$\operatorname{ord}\left(\sum\nolimits_{\text{finite}} a_i x^i\right) = n \quad \Longleftrightarrow \quad a_n \neq 0 \text{ and } a_i = 0 \text{ for all } i < n.$$

Let us adopt the convention that $\operatorname{ord} 0 > n$ for all integers $n \in \mathbf{N}$, so that

$$\operatorname{ord}(f + g) \geqslant \min(\operatorname{ord} f, \operatorname{ord} g),$$

is true in general (even if $g = -f$), and for each integer $n \geqslant 0$

$$W_n = \{f \in \Pi : \operatorname{ord} f \geqslant n\}$$

is an infinite-dimensional subspace of Π. The elements of this subspace are the polynomials divisible by x^n. The sequence formed by the subspaces W_n decreases indefinitely

$$\Pi = W_0 \supset W_1 \supset W_2 \supset \cdots \supset W_n \supset \cdots \supset \{0\}.$$

2.3.2 Existence of Bases: The Mathematical Credo

All mathematicians believe that

Every vector space $\neq \{0\}$ has a basis.

Life is indeed easier if we accept it, and the language of vector spaces is simplified if we accept it in the following general form.

Postulate. *Let E be a nonzero vector space, A_1 and A_2 two subsets of E such that A_1 is linearly independent, and A_2 generates E. Then there is a basis $B = A_1 \cup A_2'$ of E where A_2' is a subset of A_2.* ■

In particular, taking $A_2 = E$, we obtain the general incomplete basis theorem.

Corollary. Any independent subset A of a vector space $E \neq \{0\}$ may be completed into a basis of E. In particular, for any nonzero $\mathbf{x} \in E$, *there is a basis of E containing* $\mathbf{x}$. ∎

The preceding postulate can be derived from the *axiom of choice.*[2] We take it from granted. The following example is meant to illustrate the difficulties that may arise in vector spaces that are not finitely generated.

Example. Consider the vector space E of all sequences $\mathbf{a} = (a_n)_{n\geqslant 0}$ of scalars. Addition of sequences is defined componentwise

$$\mathbf{a} + \mathbf{b} = (a_n)_{n\geqslant 0} + (b_n)_{n\geqslant 0} = (a_n + b_n)_{n\geqslant 0}.$$

Multiplication of sequences by a scalar is similarly defined componentwise

$$a\mathbf{a} = a(a_n)_{n\geqslant 0} = (aa_n)_{n\geqslant 0}.$$

The particular sequences

$$\mathbf{e}_0 = (1\,0\,0\,0\,0\ldots), \quad \mathbf{e}_1 = (0\,1\,0\,0\,0\ldots), \quad \mathbf{e}_2 = (0\,0\,1\,0\,0\ldots), \quad \ldots$$

are independent. They generate the subspace consisting of the sequences which are finally 0 (think of polynomials). For example, the constant sequence

$$\mathbf{f}_0 = \mathbf{1} = (1\,1\,1\,1\,1\ldots)$$

is not in the span $\mathcal{L}(\mathbf{e}_i : i \in \mathbf{N})$ since it is *not* a *linear combination* of the $\mathbf{e}_i$ (remember that linear combinations are *finite* sums). More generally, the sequences

$$\mathbf{f}_0, \ \mathbf{f}_j = (0\,1\,2^j\,3^j\,4^j\ldots) \quad (j \geqslant 1)$$

together with the $\mathbf{e}_i$ $(i \geqslant 0)$ constitute a free subset, not a generating set. We may still add the sequences

$$(0\,1\,2^2\,3^3\,4^4\ldots), \quad (1\,1\,2!\,3!\,4!\ldots), \quad (0\,1^{1^1}\,2^{2^2}\,3^{3^3}\ldots).$$

The trouble is that we can continue forever, and never attain an explicit basis. Mathematical induction is not powerful enough to produce such a basis and a general *maximality principle* (equivalent to the axiom of choice) has to be invoked here. With it, it is easily seen that there is a basis of E containing the sequences that we have mentioned. But it is not possible to enumerate it explicitly.

---

[2]Set theory has several axiomatic foundations: In all currently adopted ones, there is such a statement. Speaking of the axiom of choice, the famous philosopher Bertrand Russel used to say: "At first it seems obvious, but the more you think about it, the stranger the deductions from this axiom seem to become; in the end you cease to understand what is meant by it!"

2.3.3 Infinite-Dimensional Examples

It is not always possible to index a basis of an infinite-dimensional vector space with the set N of natural numbers. Let us illustrate this on some examples, borrowed from algebra and analysis (we shall use complex scalars in this section, although the complex field C will only be introduced in Sec. 3.3.2 below). This section will not be referred to, and may be skipped by readers unfamiliar with the context and its methods.

Definition. A rational function $f = p/q$ *is the quotient of two polynomials* p *and* q, *where* $q \neq 0$ *is not the zero polynomial.*

We identify a polynomial p with the quotient $p/1$, and two quotients p_1/q_1 and p_2/q_2 are identified when $p_1q_2 = q_1p_2$, so that we may use simplified forms for rational functions. Each rational function has a representation $f = p/q$ in simplified form where the polynomials p and q are relatively prime (have no common factor). In this form, q is uniquely determined up to a multiplication by a nonzero constant, and the finite set Z_q of its zeros is the set of *poles* P_f of f. The value $f(x) = p(x)/q(x)$ is well-defined provided x is not in this set P_f so that f defines a map

$$D_f = \mathbf{R} - P_f \longrightarrow \mathbf{R}$$

having for domain the complement of its set of poles. Since the real (or complex) field is infinite, the common domain of two rational functions is also infinite and with the usual definition for the sum and multiplication by scalars, *the set of real (or complex) rational functions is a vector space.*

Lemma. The family $\left(\dfrac{1}{x-a}\right)_{a\in\mathbf{R}}$ *is linearly independent.*

PROOF. Consider a finite linear combination

$$\sum_{1\leqslant i\leqslant n} \frac{c_i}{x-a_i} = 0,$$

where the a_i are n distinct scalars. We have to prove that all $c_i = 0$. Without loss of generality, it is enough to show $c_1 = 0$. Consider the equality

$$\frac{c_1}{x-a_1} = -\sum_{2\leqslant i\leqslant n} \frac{c_i}{x-a_i}$$

Since the a_i ($i \geqslant 2$) are different from a_1, the sum in the right-hand side is *bounded* in a neighborhood of a_1. Hence $\frac{c_1}{x-a_1}$ has the same property: This implies $c_1 = 0$. ∎

How can we complete the preceding independent set into a basis of the whole space of rational functions? Obviously, since polynomials are rational functions (p is identified with $p/1$), we may add the basis $(x^j)_{j\geqslant 0}$ of the space of polynomials. But the linear span of the preceding functions does not contain the powers $1/(x-a)^j$ for $j \geqslant 2$. And due to the fact that the denominator of $\frac{1}{1+x^2}$ has no real root, we cannot express this rational function as a linear combination of the preceding ones. Eventually, let us indicate without proof how we may finally reach a basis of this space.

Theorem 1. *The family*

$$x^j, \quad \frac{1}{(x-a)^{j+1}}, \quad \frac{1}{(x^2+bx+c)^{j+1}}, \quad \frac{x}{(x^2+bx+c)^{j+1}}$$

$$(j \in \mathbf{N};\ a, b, c \in \mathbf{R} \text{ with } b^2 < 4c),$$

constitutes a basis of the vector space of real rational functions. ■

The meaning of this theorem is the following. Any rational function is a (finite) linear combination of simple rational functions listed in the theorem. Its interest is obvious for finding primitives of rational functions. It is also useful for computing iterated derivatives of such functions. But if we have this application in mind, we may note that while it is easy to write down the nth derivative of $\frac{1}{(x-a)^{j+1}}$, it is no so easy to compute iterated derivatives of

$$\frac{1}{(x^2+bx+c)^{j+1}}, \quad \frac{x}{(x^2+bx+c)^{j+1}}.$$

The situation is simpler if we use complex scalars since we may then factor the denominators into degree 1 polynomials.

Theorem 2. *The family*

$$x^j, \quad \frac{1}{(x-a)^{j+1}}, \quad (j \in \mathbf{N};\ a \in \mathbf{C})$$

constitutes a basis of the vector space of complex rational functions. ■

In other words, every complex rational function is a (finite) linear combination of elements of the preceding list. Grouping them suitably, we find that any complex rational function f is the sum of a polynomial $p(x)$ and of finitely many principal parts

$$\frac{c_{1,i}}{x-a_i} + \cdots + \frac{c_{m_i,i}}{(x-a_i)^{m_i}} = P_i\Big(\frac{1}{x-a_i}\Big),$$

associated to the different poles a_i of f. Here, P_i is a polynomial having zero constant term: $P_i(0) = 0$.

Finally, here is an example which plays an important part in analysis.

Theorem 3. *The family*

$$x^j e^{ax} \quad (j \in \mathbf{N};\ a \in \mathbf{R})$$

is linearly independent in the vector space of all functions $\mathbf{R} \to \mathbf{R}$.

PROOF. We have to consider (finite) linear combinations of functions of the mentioned type producing the zero function. Therefore, we introduce finitely many distinct exponents $a_i \in \mathbf{R}$, with finitely many scalars c_{ij} such that

$$\sum_{\text{finite}} c_{ij} x^j e^{a_i x} = 0 \quad \text{(vanishes identically).}$$

We must then prove that all $c_{ij} = 0$. Grouping the terms corresponding to the same exponent (with the same index i), we see that we have to prove that a finite sum $\sum p_i e^{a_i x}$ having polynomial coefficients p_i, that vanishes identically has all $p_i = 0$. We show this by induction on the number m of terms in such a sum.

(a) *Case* $m = 1$. Let $p(x)e^{ax}$ vanish identically, where p is a polynomial. Since $e^{ax}e^{-ax} = e^{ax-ax} = e^0 = 1$, the exponential never vanishes and the assumption implies that $p(x) = 0$ vanishes identically. This can only happen if $p = 0$ is the trivial polynomial (having all zero coefficients).

(b) *Induction step.* Assume that for some $m \geqslant 1$

$$\sum_{1 \leqslant i \leqslant m} p_i(x) e^{a_i x} = 0 \text{ for all } x \quad \Longrightarrow \quad p_i = 0 \quad (1 \leqslant i \leqslant m)$$

(where the p_i's are polynomials, and the a_i are distinct scalars). Consider a dependence relation having one more term

$$\sum_{1 \leqslant i \leqslant m} p_i(x) e^{a_i x} + p_{m+1}(x) e^{a_{m+1} x} = 0 \quad \text{for all } x$$

(with polynomial coefficients p_j, and a_{m+1} distinct from all preceding a_i's). If we multiply this identity by $e^{-a_{m+1}x}$, we get

$$\sum_{1 \leqslant i \leqslant m} p_i(x) e^{\bar{a}_i x} + p_{m+1}(x) = 0 \quad \text{for all } x,$$

where all $\bar{a}_i = a_i - a_{m+1}$ are distinct and *nonzero* scalars. Differentiating this identity, we infer

$$\sum_{1 \leqslant i \leqslant m} q_i(x) e^{\bar{a}_i x} + p'_{m+1}(x) = 0 \quad \text{for all } x,$$

where $q_i = \bar{a}_i p_i + p'_i$ *has the same degree as* p_i. Iterating this procedure $d + 1$ times where $d = \deg p_{m+1}$, we obtain a simpler identity

$$\sum_{1 \leqslant i \leqslant m} r_i(x) e^{\bar{a}_i x} = 0 \quad \text{for all } x,$$

still with polynomials r_i *having the same degree as* p_i. By induction assumption however, the only possibility is now $r_i = 0$ $(1 \leqslant i \leqslant m)$. The degree consideration shows that $p_i = 0$ for the same indices i. There only remains a dependence relation

$$p_{m+1}(x) e^{a_{m+1} x} = 0 \quad \text{for all } x.$$

As we have seen in the first part of the proof, it implies $p_{m+1} = 0$ also. ∎

Although we have considered *real* scalars, the reader may observe that the preceding proof also works in the *complex* case.

2.4 Appendix

2.4.1 Set Theory, Notation

A *set* is a collection of mathematical objects. It is given by a list between brackets, e.g. the set consisting of the two numbers 1 and 2 is $\{1,2\}$. For infinite sets we use dots, e.g. the set of *natural integers* is

$$\mathbf{N} = \{0,1,2,3,\ldots\}.$$

or we list the property which is characteristic of the set. For example, the set of even numbers is

$$E = \{0,2,4,6,\ldots\} = \{2n : n \text{ is a natural integer}\},$$

the set of *prime integers*, or simply the set of *primes* is

$$P = \{2,3,5,7,11,\ldots\} \subset \mathbf{N}.$$

To indicate that an element belongs to a set, we use the $\in$ symbol: Instead of "23 is a prime" we may equivalently write "$23 \in P$", which is read "23 is an element of—or belongs to—the set P of primes". The negation of $\in$ is denoted by $\notin$, e.g. $1 \notin P$: the integer 1 is not a prime.[3] The set of even integers and the set of primes are subsets of—or contained in—the set of natural numbers $\mathbf{N}$. This relation, also called *inclusion* is represented by the sign $\subset$ (as a reminder of the first letter in "contained"), e.g. $P \subset \mathbf{N}$. The notation $A \subset B$ is also symmetrically denoted by $B \supset A$. To prove an equality of two sets A and B, we may proceed by *double inclusion*, namely prove $A \subset B$ and $B \subset A$. When two sets E, F are given, we may define their *union*, denoted by $E \cup F = F \cup E$ (a reminder of the first letter in "union"): It is the set consisting of the elements which belong to at least one of the sets in question. For example, the union of the set of natural numbers $\mathbf{N}$ and the set of negative integers $\{-1,-2,-3,\ldots\}$ is the set $\mathbf{Z}$ of *rational integers* (first letter of "Zahl", which is the German word for "number"). The *intersection*, of two sets E and F, denoted by $E \cap F = F \cap E$, consists of their common elements. For example, the intersection of the set of even numbers and the set of primes is $\{2\}$, a set consisting of a single element, also called a *singleton* set. Here are two equivalent notations:

$$2 \in P \quad \text{and} \quad \{2\} \subset P.$$

If E is a set, the subsets of E constitute a new set

$$\mathcal{P}(E) = \{A : A \text{ is a subset of } E\}.$$

[3] Here is a reason for this: We like to have *unique* prime decompositions (up to order) of integers, e.g. $6 = 2 \cdot 3$. If we accepted 1 as a prime, we could write $6 = 2 \cdot 3 = 1 \cdot 2 \cdot 3 = 1 \cdot 1 \cdot 2 \cdot 3 = \ldots$, and hence would not have unique factorizations.

The *empty set* $\emptyset$ is a subset of any set E, hence $\emptyset \in \mathcal{P}(E)$ and this shows that $\mathcal{P}(E)$ is never empty! By convention, E itself is also a subset of E: $E \in \mathcal{P}(E)$. If two subsets A and B satisfy an inclusion $A \subset B$, we denote by $B - A$ the *relative complement* consisting of the elements of B not in A. The *complement* of a subset $A \subset E$ is the subset

$$A^c = E - A = \{x \in E : x \notin A\}.$$

By definition, the complement of A^c is A itself

$$(A^c)^c = A.$$

The complement of a union is the intersection of the complements

$$(A \cup B)^c = A^c \cap B^c.$$

The union of any family of sets is the set consisting of the elements which belong to at least one. For three sets, A, B, and C, this union is the set $A \cup B \cup C$. It can be obtained by first taking the union of A and B, and then the union of $A \cup B$ and C. Hence

$$(A \cup B) \cup C = A \cup B \cup C = A \cup (B \cup C).$$

This is the *associativity* of the "$\cup$" operation. Similar considerations hold for the intersection:

$$(A \cap B) \cap C = A \cap B \cap C = A \cap (B \cap C).$$

There is also a *distributivity* relation

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

(proof by double inclusion: Make a picture!). These operations make the basis of the *Boolean algebra* of subsets of a set E.

When E and F are two sets, their *Cartesian product* $E \times F$ is the set of *vertically written pairs* $\binom{x}{y}$ where x is taken in E and y in F. In particular, when $E = F$, the Cartesian product $E^2 = E \times E$ consists of pairs $\binom{x}{y}$ of elements of E, $E^3 = E \times E \times E$ consists of vertically written triples. More generally, a list of n elements of E, *written vertically*, is called an n-tuple, and

$$E^n = \underbrace{E \times E \times \cdots \times E}_{n \text{ terms}}$$

denotes the set of n-tuples of elements of E. Here are some examples of n-tuples

$$\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \in \mathbf{R}^3; \quad \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \in E^3; \quad \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in E^n \quad (x, y, z, x_i \in E).$$

We reserve the notation E_n for the set of *rows* $(x_1, x_2, \ldots, x_n)$ consisting of n elements of E.

A *map* or *mapping* from a set E to a set F is a correspondence which to *each* element $x \in E$ associates *one* element $y \in F$. We often denote a map by $f : E \to F$, and we indicate the correspondence at the level of elements by $x \mapsto y = f(x)$. For example, a map $\{1, 2, 3\} \to \mathbf{R}$ is a correspondence

$$\begin{array}{ccc} 1 & \longmapsto & x_1, \\ 2 & \longmapsto & x_2, \\ 3 & \longmapsto & x_3. \end{array}$$

Such a map may be identified with the vertically written triple $(x_i)_{i=1,2,3}$. When I is any set, a map f from I to E can also be viewed as a list of elements of E indexed by I: If $f(i) = x_i$, namely if $f : i \mapsto x_i$, we identify f to the list of its values, which constitutes a *family* $(x_i)_{i \in I}$ of elements of E, indexed by I. When j is a particular element of I, the element $x_j \in E$ is called the jth *component* of the family in question.

If $f : E \to F$, E is the *domain* and F the *target* of f, and for each $x \in E$ there is *only one* corresponding element $f(x) \in F$. The *graph* of f is the subset

$$G_f = \left\{ \binom{x}{f(x)} : x \in E \right\} \subset E \times F.$$

The *image* of f is the subset

$$\operatorname{im} f = f(E) = \{ y \in F : y = f(x) \text{ for some } x \in E \} \subset F.$$

When $f(E) = F$, we say that f is *surjective*, or *onto*. When

$$x \neq y \quad \Longrightarrow \quad f(x) \neq f(y),$$

we say that f is *injective*, or 1-1 (read "one-to-one"). When both conditions hold, we say that f is *bijective*, or 1-1 onto. When there is a bijection between two sets E and F, they are *equipotent*, and this is a definition of the fact that they have the *same cardinality* (same number of elements). Let $f : E \to F$ be a map, and A a subset of E. Restricting the domain of f to A, we obtain a map $f|_A : A \to F$, called restriction of f to the subset A. The numerical map on E defined by

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A, \end{cases}$$

is called *characteristic function* of A.

Fundamental Sets of Numbers

We have introduced the canonical sets $\mathbf{N}$ and $\mathbf{Z}$. Here are larger ones

- $\mathbf{Q}$: set of rational numbers $m/n \quad (m, n \in \mathbf{Z}, n \neq 0)$,
- $\mathbf{R}$: set of real numbers (scalars),
- $\mathbf{C}$: set of complex numbers (to be introduced in Sec. 3.3.2).

About the "sigma" summation symbol

To save space, we often use the "sigma" summation symbol as follows

$$f(1)+f(2)+\cdots+f(n)=\sum_{1\leqslant i\leqslant n} f(i).$$

Here, the *free variable* i may be replaced by any other:

$$\sum_{1\leqslant i\leqslant n} f(i)=\sum_{1\leqslant j\leqslant n} f(j)=\sum_{1\leqslant \ell\leqslant n} f(\ell).$$

The result depends on the choice of function f *and* on the length n of the summation. The *triangle inequality*

$$|a_1+\cdots+a_n|\leqslant |a_1|+\cdots+|a_n|$$

is now simply written

$$\Big|\sum_{1\leqslant i\leqslant n} a_i\Big|\leqslant \sum_{1\leqslant i\leqslant n} |a_i|.$$

By convention, a sum containing no term is 0: $\sum_{0<n<1} f(n)=0$ since there are no integers between 0 and 1. A summation is extended over a certain range, determined by a property of its free variable, explicitly described under the Σ symbol. In linear algebra, the free variable is often an *index*, as in

$$a_1+\cdots+a_n=\sum_{1\leqslant i\leqslant n} a_i.$$

An nth degree polynomial is simply written

$$a_n x^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0=\sum_{0\leqslant i\leqslant n} a_i x^i.$$

A *product* of terms may similarly be represented by a big Π symbol

$$f(1)f(2)\cdots f(n)=\prod_{1\leqslant i\leqslant n} f(i)=\prod_{1\leqslant j\leqslant n} f(j).$$

By convention, a product with no term is equal to 1.

Logical symbols

Mathematicians use the *logical quantifiers* as abbreviations:

$\forall\, x\ldots$ is read "for all $x\ldots$", and $\exists\, x\ldots$ is read "there exists an $x\ldots$"

2.4.2 Axioms for Fields of Scalars

The scalars, elements of a field K, are represented by $a, b, c, \ldots$

$$\left.\begin{array}{ll}
1. & a+(b+c)=(a+b)+c \\
2. & a+b=b+a \\
3. & \exists\, 0 \in K, \forall\, a \in K:\ 0+a=a \\
4. & \forall\, a \in K,\ \exists\, -a \in K:\ a+(-a)=0
\end{array}\right\}\ \textit{K is an additive Abelian group}$$

$$\left.\begin{array}{ll}
5. & a(bc)=(ab)c \\
6. & ab=ba \\
7. & \exists\, 1\,(\neq 0) \in K, \forall\, a \in K:\ 1a=a \\
8. & \forall\, a \in K,\ a \neq 0,\ \exists\, a^{-1} \in K: aa^{-1}=1
\end{array}\right\}\ \textit{K}-\{0\}\ \textit{is a multiplicative Abelian group}$$

9. $a(b+c)=ab+ac$ *(Distributivity)*

2.5 Exercises

1. Show that the columns containing pivots of a row-reduced array are independent.

2. Let **a**, **b**, **c** be three elements of a vector space. (*a*) If **a** and **b** are independent, **b** and **c** are independent, does it follow that **a** and **c** are independent? (*b*) If **a** and **b** are independent, **b** and **c** are independent, **a** and **c** are independent, does it follow that **a**, **b**, and **c** are independent?

3. By definition, the monomials $1,\ x,\ x^2, \ldots,\ x^n, \ldots$ form a basis of the space of polynomials Π. Show that the polynomials $1,\ x-1,\ (x-1)^2, \ldots,\ (x-1)^n, \ldots$ also form a basis of this space.

4. (*a*) The set of functions $t \mapsto f(t) = a\cos(t+b)$, where a and b are arbitrary real numbers, is a natural vector space: Explain why.
(*b*) What is the dimension of the vector space generated by

$$\cos^2 t,\ \cos 2t,\ \sin^2 t,\ \text{and}\ \sin 2t?$$

Is this vector space the same as the space generated by 1, $\cos 2t$, and $\sin 2t$?

5. The set of linear equations $a_1x_1 + \cdots + a_nx_n = b$ is a vector space E. What is its dimension? Is the subset consisting in equations having a solution a vector subspace of E?

6. Show that the intersection of two subspaces of a vector space E is a vector subspace of E. More generally, show that the intersection of any family of subspaces of a vector space E is a vector subspace of E. Give an example of two subspaces V, W of $\mathbf{R}^2$ such that $V \cup W$ is not a subspace. Show that a space of dimension greater than or equal to 2 (over the real field $\mathbf{R}$) is not a union of finitely many 1-dimensional subspaces.

7. What is the linear span of the 2^n vectors $\pm\mathbf{e}_1 \pm \mathbf{e}_2 \pm \cdots \pm \mathbf{e}_n$ in $\mathbf{R}^n$? Extract a maximal subset of independent vectors from the preceding family.

8. Find a basis of the subspace of $\mathbf{R}^n$ consisting of the n-tuples (x_i) such that $\sum_{1 \leqslant i \leqslant n} x_i = 0$.

9. The set of sequences $(a_i)_{i \geqslant 0} = (a_0, a_1, \ldots, a_n, \ldots)$ of real numbers is a vector space E if we define scalar multiplication and addition componentwise

$$\begin{aligned} c(a_0, a_1, \ldots, a_n, \ldots) &= (ca_0, ca_1, \ldots, ca_n, \ldots), \\ (a_0, a_1, \ldots, a_n, \ldots) + (b_0, b_1, \ldots, b_n, \ldots) &= (a_0 + b_0, a_1 + b_1, \ldots, a_n + b_n, \ldots). \end{aligned}$$

Which of the following conditions characterize a vector subspace of E:

(a) $a_{i+1} = a_i^2 \quad (i \geqslant 0)$

(b) $a_{i+1} = 2a_i + 1 \quad (i \geqslant 0)$

(c) $a_{i+1} = a_i + a_{i-1} \quad (i \geqslant 1)$

(d) $a_i \geqslant 0 \quad (i \geqslant 0)$

(e) a_0 is an integer

(f) The a_i are 0 for i sufficiently large.

10. The *sum* of two vector subspaces V and W of a fixed vector space E is *defined* by

$$V + W = \{\mathbf{v} + \mathbf{w} : \mathbf{v} \in V, \mathbf{w} \in W\} \subset E.$$

Show that $V+W$ is a vector subspace of E. What is $V+V$? When is $V+W = V$? What are the properties of this composition law: Associativity, commutativity, existence of a neutral element? Let us define an external multiplication of subspaces by scalars as follows

$$0V = \{0\}, \quad \text{and} \quad aV = V \text{ if } a \neq 0.$$

Check the formal properties satisfied by this external multiplication: Is the set of subspaces of E a vector space with respect to these addition and multiplication?

11. (a) Let T be a fixed positive number. Prove that the set of T-periodic functions $f : \mathbf{R} \to \mathbf{R}$ is a vector space.
(b) Is the set of periodic functions (arbitrary periods) a vector space?

12. Let Π be the space of polynomials in one variable x. (a) The set of polynomials having degree greater than or equal to some positive integer n is not a subspace of the vector space of polynomials. Why?

(*b*) Let V be a subspace of a vector space E. Is the complement $V^c = \{\mathbf{v} \in E : \mathbf{v} \notin V\}$ a subspace of E? Do you see a relation between this question and the preceding one?
(*c*) Prove that the subset of polynomials vanishing at the point $x = 2$ is a subspace of Π. Is the subset of polynomials divisible by a fixed positive power $(x-2)^d$ of $x-2$ a subspace of Π? If yes, give a basis of this subspace. For $m \geqslant 1$, let V_m be the subspace spanned by $1, x^m, x^{2m}, \ldots, x^{km}, \ldots$ Give a condition on m and ℓ in order to have V_m contained in V_ℓ.

13. (*a*) Show that for any field K and any set E, the set of functions from E to K

$$\mathcal{F}(E; K) = \{f : E \to K\}$$

is a vector space over K.
(*b*) Check that the set $\mathbf{F}_2 = \{0, 1\}$ with addition and multiplication defined by

$$\begin{array}{c|cc} + & 0 & 1 \\ \hline 0 & 0 & 1 \\ 1 & 1 & 0 \end{array} \qquad \begin{array}{c|cc} \cdot & 0 & 1 \\ \hline 0 & 0 & 0 \\ 1 & 0 & 1 \end{array}$$

is a field.
(*c*) For any subset $A \subset E$ let $\mathbb{1}_A$ denote the characteristic function $E \to \mathbf{F}_2$, defined by

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A, \end{cases}$$

so that the map $A \mapsto \mathbb{1}_A$ defines a bijection $\mathcal{P}(E) \xrightarrow{\sim} \mathcal{F}(E; \mathbf{F}_2)$. The power set of E can thus be identified with a vector space. To which operations on subsets do addition and multiplication of functions correspond?
(*d*) The field $\mathbf{F}_2$ defined in (*b*) *is* the field of integers mod 2: 0 represents the "*even*" class, and 1 the "*odd*" class. Check that the field of integers mod 3 has addition and multiplication tables

$$\begin{array}{c|ccc} + & 0 & 1 & 2 \\ \hline 0 & 0 & 1 & 2 \\ 1 & 1 & 2 & 0 \\ 2 & 2 & 0 & 1 \end{array} \qquad \begin{array}{c|ccc} \cdot & 0 & 1 & 2 \\ \hline 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 2 \\ 2 & 0 & 2 & 1 \end{array}$$

In fact, for any prime integer $p = 2, 3, 5, 7, 11, 13, \ldots$, the classes mod p constitute a finite field $\mathbf{F}_p$ with p elements. In such a field, the scalars

$$1,\ 1+1,\ 1+1+1,\ \ldots,\ n1 = \underbrace{1+1+\cdots+1}_{n \text{ terms}},\ \ldots$$

are not all different!

Notes

Here is the feeling of *dimension*, as vulgarized by Karl Sabbagh

> (...) *moving from two dimensions to three, three to four, and four to many carries with it an idea of a multidimensional "space" in which things exist or happen in the same way as in our familiar three-dimensional space. If I hold my finger in the air above my desktop, I can describe its position with three numbers: the perpendicular distance from the surface and the horizontal distances from the two edges of the desk, which are at right angles. I don't know what it would mean to describe it by four figures, or five or six...*
>
> [from *The Riemann Hypothesis*]

Repeat definitions until you understand them!

Chapter 3

Matrix Multiplication

In typography, the word *matrix* refers to a lead rectangle on which the characters are placed in view of the printing process. By analogy, in mathematics the term *matrix* refers to any rectangular array containing mathematical entities. If it has m rows and n columns, it is a *matrix* of *size* $m \times n$ (read m by n). In Chapter 1, arrays were viewed as representations of linear systems. Here, we deal with matrices in a more general and abstract way: Matrices of the same size form a vector space. Two matrices of suitable size may even be *multiplied* together.

3.1 Row by Column Multiplication

3.1.1 Linear Fractional Transformations

To guess how to define the multiplication of 2×2 matrices, let us consider the composition of *linear fractional transformations*. Such a function is defined by

$$y = f(x) = \frac{ax+b}{cx+d} \quad \text{where } ad \neq bc.$$

The condition $ad \neq bc$ implies that c and d are not both 0, hence the denominator is not identically 0. Moreover it also shows $\frac{a}{c} \neq \frac{b}{d}$ so that numerator and denominator are not proportional. The function f is not a constant, and is well-defined except where the denominator vanishes. When $c \neq 0$, the point $x = -d/c$ is not in the domain of f: This is a vertical asymptote for its graph. (For $y = a/c$, there is a horizontal asymptote; linear fractional transformations are bijective maps if we add a point "infinity" to the line, and extend in a natural way the definition of f.) If we make a change of variable

$$x = g(t) = \frac{\alpha t+\beta}{\gamma t+\delta} \quad \text{where } \alpha\delta \neq \beta\gamma,$$

we find a composite in the form of another linear fractional transformation

$$y = f(g(t)) = \frac{a\frac{\alpha t+\beta}{\gamma t+\delta} + b}{c\frac{\alpha t+\beta}{\gamma t+\delta} + d} = \frac{(a\alpha + b\gamma)t + (a\beta + b\delta)}{(c\alpha + d\gamma)t + (c\beta + d\delta)}.$$

This suggests to *define* the product of 2×2 matrices by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} = \begin{pmatrix} a\alpha + b\gamma & a\beta + b\delta \\ c\alpha + d\gamma & c\beta + d\delta \end{pmatrix}.$$

3.1.2 Linear Changes of Variables

The multiplication law for matrices (to be introduced below) can be motivated more generally as follows. In

$$\begin{cases} y_1 = & a_{11}x_1 + \cdots + a_{1n}x_n \\ \vdots & \vdots \\ y_m = & a_{m1}x_1 + \cdots + a_{mn}x_n, \end{cases}$$

let us consider a linear change of variables

$$\begin{cases} x_1 = & b_{11}t_1 + \cdots + b_{1p}t_p \\ \vdots & \vdots \\ x_n = & b_{n1}t_1 + \cdots + b_{np}t_p. \end{cases}$$

For the ith variable y_i we have

$$\begin{aligned} y_i &= a_{i1}x_1 + \quad \cdots \quad + a_{in}x_n \\ &= a_{i1}(b_{11}t_1 + \quad \cdots \quad + b_{1p}t_p) \\ &\qquad\quad + \quad \cdots \quad + \\ &\quad\ a_{in}(b_{n1}t_1 + \quad \cdots \quad + b_{np}t_p) \\ &= \left(\textstyle\sum_k a_{ik}b_{k1}\right)t_1 + \quad \cdots \quad + \left(\textstyle\sum_k a_{ik}b_{kp}\right)t_p. \end{aligned}$$

Hence we have found the expressions

$$\begin{cases} y_1 = & c_{11}t_1 + \cdots + c_{1p}t_p \\ \vdots & \vdots \\ y_m = & c_{m1}t_1 + \cdots + c_{mp}t_p, \end{cases}$$

with the coefficients

$$c_{ij} = \sum_{1 \leqslant k \leqslant n} a_{ik}b_{kj} \quad (1 \leqslant i \leqslant m, 1 \leqslant j \leqslant p).$$

In short, the result of the linear change of variables is

$$y_i = \textstyle\sum_j c_{ij} t_j \quad \text{where} \quad c_{ij} = \textstyle\sum_k a_{ik} b_{kj}.$$

In particular when $m = n = p = 2$ we recover the result of the preceding subsection, where the notation was

$$a_{11} = a,\ a_{12} = b, \ldots \qquad b_{11} = \alpha,\ b_{12} = \beta, \ldots$$

3.1.3 Definition of the Matrix Product

The product of a matrix A of size $m \times n$ and a matrix B of size $n \times p$ is the matrix $C = AB$ of size $m \times p$ having the entries

$$\begin{aligned} c_{ij} &= a_{i1} b_{1j} + a_{i2} b_{2j} + \cdots + a_{in} b_{nj} \\ &= \sum_{1 \leqslant k \leqslant n} a_{ik} b_{kj} \quad (1 \leqslant i \leqslant m, 1 \leqslant j \leqslant p). \end{aligned}$$

Observe that the entry c_{ij} can be computed from the knowledge of the ith row of A and the jth column of B only. The coefficients of the product are obtained by multiplication of the *rows* of A by the *columns* of B. Here is an illustration of the coefficients involved in the computation of c_{23}

$$\begin{pmatrix} \circ & \circ & \circ & \circ & \circ \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \circ & \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ & \circ \end{pmatrix} \begin{pmatrix} \circ & \circ & \bullet & \circ \\ \circ & \circ & \bullet & \circ \\ \circ & \circ & \bullet & \circ \\ \circ & \circ & \bullet & \circ \\ \circ & \circ & \bullet & \circ \end{pmatrix} = \begin{pmatrix} \circ & \circ & \circ & \circ \\ \circ & \circ & \blacksquare & \circ \\ \circ & \circ & \circ & \circ \\ \circ & \circ & \circ & \circ \end{pmatrix}.$$

Taking $p = 1$, we see that a matrix A of size $m \times n$ can be multiplied by an n-tuple $\mathbf{x}$ and the result is the m-tuple

$$A\mathbf{x} = \begin{pmatrix} a_{11}x_1 & + & \cdots & + & a_{1n}x_n \\ a_{21}x_1 & + & \cdots & + & a_{2n}x_n \\ \vdots & & & & \vdots \\ a_{m1}x_1 & + & \cdots & + & a_{mn}x_n \end{pmatrix}.$$

Hence the simple equation $A\mathbf{x} = 0$ is equivalent to the homogeneous system

$$(HS) \qquad \begin{cases} a_{11}x_1 & + & \cdots & + & a_{1n}x_n & = 0 \\ a_{21}x_1 & + & \cdots & + & a_{2n}x_n & = 0 \\ \vdots & & & & \vdots & \vdots \\ a_{m1}x_1 & + & \cdots & + & a_{mn}x_n & = 0. \end{cases}$$

The symbolic way of writing linear systems is now explained by matrix multiplication: The general linear system is similarly written $A\mathbf{x} = \mathbf{b}$.

Example: Discretization of Graphs

The screen of a computer is a matrix of pixels 0 and 1 (black and white case). Any picture can thus be identified with a matrix of 0's and 1's. In particular, the graph of a function is such a matrix.

Let us assume that our screen is a square, and let $A = (a_{ij})$ be the digitalization of the graph of a function f. By definition we only take one active pixel in each column. In the neighborhood of a point where the derivative is large, this is not very good on the visual point of view; but this respects the fundamental mathematical principle that each value of x has only one image $f(x)$. Hence

$$a_{ij} = \begin{cases} 1 & \text{if} \quad i = f(j) \\ 0 & \text{otherwise.} \end{cases}$$

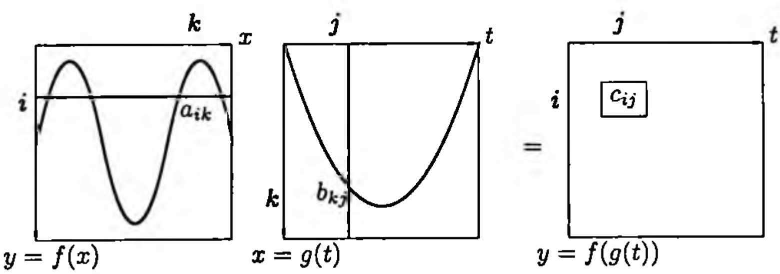
In a picture of this graph, the row number increases towards the bottom, contrary to the usual convention for Cartesian coordinates: We respect the matrix convention. But the j-axis has the usual orientation. Consider a second digitalized graph $B = (b_{ij})$:

$$b_{ij} = \begin{cases} 1 & \text{if} \quad i = g(j) \\ 0 & \text{otherwise.} \end{cases}$$

The matrix product $C = AB = (c_{ij})$ has the coefficients

$$c_{ij} = \sum_k a_{ik} b_{kj}.$$

In such a product, there is only one nonzero element in the jth column of B (but there may be no, or many nonzero elements in the ith row of A).



In the row by column product, a single b coefficient is nonzero, the one for which $k = g(j)$, and this one is $b_{kj} = 1$. Hence

$$c_{ij} = a_{ig(j)}.$$

This coefficient is often zero: It is nonzero precisely when its first index i is $f(k)$

$$c_{ij} = a_{ig(j)} = \begin{cases} 1 & \text{if} \quad i = f(g(j)) \\ 0 & \text{otherwise.} \end{cases}$$

We recognize the matrix of the digitalized graph of the composite $f \circ g$:

The matrix product of digitalized graphs is the digitalized graph of the composition.

Properties of the Matrix Product

By definition, the first row of a matrix product is obtained by products of the first row of A successively with all the columns of B

$$\begin{pmatrix} \boxed{a_{11}\, a_{12} \dots a_{1n}} \\ \dots \\ \dots \end{pmatrix} \begin{pmatrix} \vdots & \boxed{\begin{matrix} b_{1j} \\ b_{2j} \\ \dots \\ b_{nj} \end{matrix}} & \vdots & \vdots \end{pmatrix} = \begin{pmatrix} \vdots & \boxed{\Sigma_k a_{1k} b_{kj}} & \vdots & \vdots \\ & \dots & & \\ & \dots & & \end{pmatrix}.$$

One crucial case of product is when A is a row of length n and B a column of the same length: The product is a 1×1 matrix, namely a scalar (no need to embed it between parentheses)

$$(a_1, a_2, \dots, a_n) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = a_1 x_1 + a_2 x_2 + \dots + a_n x_n.$$

Conversely, the product of a unicolumn matrix by a row matrix leads to a rectangular matrix

$$\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} (a_1, a_2, \dots, a_n) = \begin{pmatrix} b_1 a_1 & b_1 a_2 & \dots & b_1 a_n \\ b_2 a_1 & b_2 a_2 & \dots & b_2 a_n \\ \vdots & \vdots & & \vdots \\ b_m a_1 & b_m a_2 & \dots & b_m a_n \end{pmatrix}.$$

The matrix product is not commutative: Even for square matrices, we may find $AB \neq BA$. However, it is obviously *distributive* with respect to addition. For example, the square of a sum can be computed as follows

$$\begin{aligned} (A+B)^2 &= (A+B)(A+B) = A(A+B) + B(A+B) \\ &= A^2 + AB + BA + B^2. \end{aligned}$$

But here, $AB + BA$ is not usually equal to $2AB$.

Here is another interesting product:

$$\begin{pmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1n} \\ a_{21} & \dots & a_{2j} & \dots & a_{2n} \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ a_{m1} & \dots & a_{mj} & \dots & a_{mn} \end{pmatrix} \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{jj} \\ \vdots \\ a_{jm} \end{pmatrix}.$$

We infer that the matrix product by $\mathbf{e}_j$ *extracts the jth column* of the matrix A

$$A\,\mathbf{e}_j = \mathbf{a}_j = j\text{th column of } A.$$

From this, we may deduce the next important result.

Theorem. *The matrix product is associative: $(AB)C = A(BC)$ for matrices of adapted sizes.*

PROOF. Let $\mathbf{b}_j$ denote the jth column of B. By definition of the multiplication "row by column", the jth column of AB is $A\mathbf{b}_j$. On the other hand, we have seen that this column is also $(AB)(\mathbf{e}_j)$. This shows

$$A(B\mathbf{e}_j) = A\mathbf{b}_j = (j\text{th column of } AB) = (AB)(\mathbf{e}_j).$$

For any column vector $\mathbf{x} = \sum x_i\mathbf{e}_i$ we easily deduce

$$A(B\mathbf{x}) = (AB)\mathbf{x}.$$

Namely

$$j\text{th column of } (AB)C \text{ is } (AB)\,(j\text{th column of } C) = (AB)\mathbf{c}_j,$$

while the first part proves

$$j\text{th column of } A(BC) \text{ is } A\,(j\text{th column of } (BC)) = A(B\mathbf{c}_j).$$

The theorem is proved. ■

Definition. *A* diagonal matrix *is a square matrix having zero entries outside its main diagonal: $a_{ij} = 0$ for all $i \neq j$.*

The products with a diagonal matrix are especially simple to compute. Here is an example

$$\begin{pmatrix} c_1 & 0 & 0 & \dots & 0 \\ 0 & c_2 & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ 0 & \dots & 0 & \dots & c_n \end{pmatrix} \begin{pmatrix} a_{11} & \dots & a_{1j} & \dots & a_{1n} \\ a_{21} & \dots & a_{2j} & \dots & a_{2n} \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ a_{n1} & \dots & a_{nj} & \dots & a_{nn} \end{pmatrix}$$

$$= \begin{pmatrix} c_1a_{11} & \dots & c_1a_{1j} & \dots & c_1a_{1n} \\ c_2a_{21} & \dots & c_2a_{2j} & \dots & c_2a_{2n} \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ c_na_{n1} & \dots & c_na_{nj} & \dots & c_na_{nn} \end{pmatrix}.$$

A diagonal matrix having diagonal entries $c_1, \dots, c_n$ (as in the preceding equality) may safely be abbreviated by $\mathrm{diag}\,(c_1, \dots, c_n)$.

Definition. *The identity matrix I_n is the diagonal matrix*

$$I_n = \mathrm{diag}\,(1, \dots, 1) \quad (\textit{size } n \times n)$$

having 1's on its main diagonal.

The conventional notation is $I_n = (\delta_{ij})$ where

$$\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j, \end{cases}$$

is the *Kronecker symbol*. Hence

$$I_n = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{pmatrix}.$$

These matrices act as multiplicative units: If A has size $m \times n$, then

$$I_m A = A = A I_n.$$

Definition. *A matrix A of size $m \times n$ is* left invertible *when there is a matrix B of size $n \times m$ such that $BA = I_n$.*

When $BA = I_n$, we say that B is a *left inverse* of A. One uses a similar terminology for "right" instead of "left": A *right inverse* C of A is characterized by $AC = I_m$. The case of square matrices is particularly important.

Definition. *A square matrix A is* invertible *when there is a square matrix B of the same size which is both a left and a right inverse of A, namely when*

$$AB = BA = I_n.$$

3.1.4 The Map Produced by Matrix Multiplication

Each matrix A of size $m \times n$ defines a map

$$\begin{array}{rccc} L_A: & \mathbf{R}^n & \longrightarrow & \mathbf{R}^m \\ & \mathbf{x} & \longmapsto & \mathbf{y} = A\mathbf{x}. \end{array}$$

In fact, the matrix product has been defined so as to reproduce the composition of maps. Let us see this on the example of 2×2 matrices. Using the same notation as at the beginning of this chapter, we consider the transformations

$$\mathbf{y} = A\mathbf{x}, \quad \mathbf{x} = B\mathbf{t},$$

or in extenso

$$\begin{cases} y_1 = ax_1 + bx_2 \\ y_2 = cx_1 + dx_2, \end{cases} \qquad \begin{cases} x_1 = \alpha t_1 + \beta t_2 \\ x_2 = \gamma t_1 + \delta t_2. \end{cases}$$

Hence

$$\begin{cases} y_1 = a(\alpha t_1 + \beta t_2) + b(\gamma t_1 + \delta t_2) = (a\alpha + b\gamma)t_1 + (a\beta + b\delta)t_2 \\ y_2 = c(\alpha t_1 + \beta t_2) + d(\gamma t_1 + \delta t_2) = (c\alpha + d\gamma)t_1 + (c\beta + d\delta)t_2. \end{cases}$$

As with linear fractional transformations, the matrix product has been defined in such a way as to have

$$\mathbf{y} = A\mathbf{x} = A(B\mathbf{t}) \quad \text{equal to} \quad \mathbf{y} = (AB)\mathbf{t}.$$

This equality

$$(AB)\mathbf{t} = A(B\mathbf{t})$$

shows that the mappings

$$\mathbf{t} \overset{L_{AB}}{\longmapsto} \mathbf{y} = (AB)(\mathbf{t}),$$

$$\mathbf{t} \overset{L_B}{\longmapsto} \mathbf{x} = B\mathbf{t} \overset{L_A}{\longmapsto} \mathbf{y} = A\mathbf{x} = A(B\mathbf{t})$$

are equal:

Matrix multiplication corresponds to composition of maps

$$\mathbf{t} \to \boxed{\boxed{\;L_B\;} \;\underset{L_{AB}}{\to}\; \boxed{\;L_A\;}} \to L_A(L_B(\mathbf{t})).$$

In fact, the matrix product is precisely defined in such a way as to have this property. Since the composition of mappings is associative, it is then obvious that the matrix product is also associative.

A consequence of the equalities

$$A\,\mathbf{e}_j = \mathbf{a}_j = j\text{th column of } A$$

is the following. The image of the canonical basis by the mapping L_A is the family of columns of A. The image of the whole space $\mathbf{R}^n$ by L_A consists of the linear combinations of the columns of A, hence is the *column space of A*

$$\operatorname{im}(L_A) = \mathcal{L}(\text{columns of } A).$$

3.2 Row Operations and Matrix Multiplication

With respect to coefficientwise addition and multiplication by scalars, the matrices of fixed size $m \times n$ form a vector space of dimension mn. The canonical basis of this space is

$$(E_{ij})_{1\leqslant i\leqslant m,\, 1\leqslant j\leqslant n},$$

where E_{ij} denotes the matrix having only one nonzero coefficient—a 1—placed at the intersection of the ith row and jth column.

3.2.1 Elementary Matrices

It is easy to formulate the row operations as matrix products.

Definition. *The* elementary matrices *are the matrices obtained from an identity matrix I_m by a single row operation.*

The identity matrix I_m itself is an elementary matrix: It is obtained from I_m by multiplying any of its rows by the nonzero scalar 1. Let us now consider the three nontrivial cases separately.

(1) Let $E_i(c)$ denote the elementary matrix obtained from I_m by multiplying its ith row by the nonzero scalar c. Hence

$$E_i(c) = I_m + (c-1)E_{ii}.$$

Then, for any matrix A of size $m \times n$, $E_i(c)A$ is obtained from A by *multiplying its ith row by c*. For example

$$\begin{pmatrix} c & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a_1 & \dots & a_n \\ b_1 & \dots & b_n \end{pmatrix} = \begin{pmatrix} ca_1 & \dots & ca_n \\ b_1 & \dots & b_n \end{pmatrix}.$$

If $c \neq 0$, this is a row operation. Observe that abbreviating the rows by ρ_i, the preceding matrix product can simply be written

$$\begin{pmatrix} c & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix} = \begin{pmatrix} c\rho_1 \\ \rho_2 \end{pmatrix},$$

as if the rows were scalars. The elementary matrices of the first type $E_i(c)$ are diagonal matrices: They have zero entries a_{ij} $(i \neq j)$ outside the main diagonal.

(2) Next, consider the elementary matrix obtained from I_m by exchange of its ith and jth rows

$$P_{ij} = P_{ji} = I_m - E_{ii} + E_{ij} - E_{jj} + E_{ji}.$$

A left multiplication by P_{ij} *exchanges the ith and jth rows of A*. For example

$$P_{12} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 0 & & \\ 0 & 0 & 1 & & \vdots \\ \vdots & & & \ddots & 0 \\ 0 & & \dots & 0 & 1 \end{pmatrix}.$$

Here is a typical case

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} a_1 & \dots & a_n \\ b_1 & \dots & b_n \end{pmatrix} = \begin{pmatrix} b_1 & \dots & b_n \\ a_1 & \dots & a_n \end{pmatrix},$$

which can be abbreviated

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix} = \begin{pmatrix} \rho_2 \\ \rho_1 \end{pmatrix}.$$

(3) Finally, let $E_{ij}(c)$ be the matrix obtained by adding c times the jth row to the ith row of I_m

$$E_{ij}(c) = I_m + cE_{ij} \quad (i \neq j).$$

This matrix has a c located in the (ij) position, and when $i > j$, this coefficient appears *below* the main diagonal. A left multiplication by $E_{ij}(c)$ $(i \neq j)$, *replaces the ith row ρ_i of A by $\rho_i + c\rho_j$*. For example, if we add c times the first row to the second row of $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, we obtain $\begin{pmatrix} 1 & 0 \\ c & 1 \end{pmatrix}$, and a left multiplication by this matrix replaces the second row ρ_2 of A by $\rho_2 + c\rho_1$

$$\begin{pmatrix} 1 & 0 \\ c & 1 \end{pmatrix} \begin{pmatrix} a_1 & \dots & a_n \\ b_1 & \dots & b_n \end{pmatrix} = \begin{pmatrix} a_1 & \dots & a_n \\ ca_1 + b_1 & \dots & ca_n + b_n \end{pmatrix},$$

or simply

$$\begin{pmatrix} 1 & 0 \\ c & 1 \end{pmatrix} \begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix} = \begin{pmatrix} \rho_1 \\ c\rho_1 + \rho_2 \end{pmatrix}.$$

Let us summarize our observations:

> *A left multiplication by an elementary matrix reproduces the row operation that produced it.*

Caution. Some books reserve the denomination of *elementary matrix* to the third type $E_{ij}(c)$ $(i \neq j)$. Products of these special elementary matrices cannot produce matrices of the first and second general type (see Sec. 10.3.2). However, consider the following sequence of row operations, where c is a nonzero scalar:

$$\begin{pmatrix} \rho_1 \\ \rho_2 \end{pmatrix} \sim \begin{pmatrix} \rho_1 \\ c\rho_1 + \rho_2 \end{pmatrix} \sim \begin{pmatrix} -c^{-1}\rho_2 \\ c\rho_1 + \rho_2 \end{pmatrix} \sim \begin{pmatrix} -c^{-1}\rho_2 \\ c\rho_1 \end{pmatrix}.$$

It corresponds to the identity

$$\begin{pmatrix} 1 & 0 \\ c & 1 \end{pmatrix} \begin{pmatrix} 1 & -c^{-1} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ c & 1 \end{pmatrix} = \begin{pmatrix} 0 & -c^{-1} \\ c & 0 \end{pmatrix}.$$

In particular, up to a sign, a permutation of rows may be obtained using only the special operations of the third type

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

Since row operations are invertible, the elementary matrices are left invertible, and in fact invertible: The inverses are elementary matrices of the same form

$$E_i(c)^{-1} = E_i(1/c), \quad P_{ij}^{-1} = P_{ij}, \quad E_{ij}(c)^{-1} = E_{ij}(-c).$$

Proposition. *All products of elementary matrices are invertible.*

PROOF. If the inverse of the elementary matrix E_i is F_i ($1 \leqslant i \leqslant k$), then the inverse of $E_1 \cdots E_k$ is $F_k \cdots F_1$ as the following computations show

$$\begin{aligned} E_1 \cdots \underbrace{E_k F_k}_{=I} \cdots F_1 &= \cdots = E_1 F_1 = I, \\ F_k \cdots \underbrace{F_1 E_1}_{=I} \cdots E_k &= \cdots = F_k E_k = I \end{aligned}$$

(undo first what has been done last, see Sec. 4.5.1). ∎

Alternatively, one may argue as follows. Let A be left invertible, say $BA = I$, and assume that *this left inverse is also left invertible*, say $CB = I$. (In our case, B is a product of elementary matrices, hence left invertible.) I claim that A is invertible, with inverse B. Observe that by associativity of the matrix product

$$CBA = \begin{cases} (CB)A = IA = A \\ C(BA) = CI = C. \end{cases}$$

Hence $C = A$, and $CB = I$ is simply $AB = I$. This proves that B is also a right inverse of A.

3.2.2 An Inversion Algorithm

Here is a useful definition.

Definition. *A square matrix $A = (a_{ij})$ is* upper-triangular *when it has zero entries below its main diagonal: $a_{ij} = 0$ for all $i > j$. It is* lower-triangular *when it has zero entries above its main diagonal: $a_{ij} = 0$ for all $i < j$. It is* triangular *when it is either upper-triangular or lower-triangular.*

For example, if we add a multiple of the ith row to the jth row of the identity matrix, we obtain an upper-triangular matrix when $i > j$, and a lower-triangular one when $i < j$. A row-reduced form of a square matrix is a particular case of an upper-triangular matrix, but the following upper-triangular matrix is not row-reduced

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Proposition. *Any square matrix A of* maximal rank *is invertible.*

PROOF. Take a square matrix A of maximal rank $r = m = n$. Using suitable row operations, we may bring A into echelon form, say $A \sim U$. Since by assumption A has maximal rank, this echelon form has 1's on the diagonal.

Using the matrix multiplication formalism for the row operations we get

$$U = E_k \cdots E_1 A = \begin{pmatrix} 1 & * & \dots & * \\ 0 & 1 & * & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Pursuing suitable row operations, we may attain a row-reduced echelon form, where 0's are to be found above each pivot: The only possibility for this row-reduced echelon form is the identity matrix

$$V = E_\ell \cdots E_k \cdots E_1 A = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \vdots \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & 1 \end{pmatrix} = I_n.$$

Hence after multiplication by a finite number of elementary matrices, we reach the identity matrix

$$E_\ell \cdots E_1 A = I_n.$$

This shows that the matrix $B = E_\ell \cdots E_1$ is a left inverse of A. But we have proved that B is a product of elementary matrices, hence is invertible:

$$B^{-1} = B^{-1} I = B^{-1}(BA) = (B^{-1}B)A = IA = A.$$

The inverse of A is $B = E_\ell \cdots E_1$. ■

Algorithm. To find the inverse of a square matrix A of maximal rank, we have to perform row operations on A until we reach the *row-reduced echelon form* $\mathrm{rref}(A) = I$. The inverse A^{-1} is the product of the elementary matrices used to achieve this form. Here is a practical way of proceeding. Observe that if we carry the same row operations on the extended array $(A \mid I)$, of size $n \times 2n$, each such operation acts on the second part as well, namely

$$E(A \mid I) = (EA \mid EI) = (EA \mid E).$$

The second part of the array acts as a memory for the operations performed. Hence finally

$$E_\ell \cdots E_1 (A \mid I) = (E_\ell \cdots E_1 A \mid E_\ell \cdots E_1) = (I \mid A^{-1}).$$

Here is a procedure for inverting a square matrix A of maximal rank:

> *The inverse of A is found in the second half of the reduced row-echelon form of $(A \mid I)$.*

Example. Let us find the inverse of the matrix

$$A = \begin{pmatrix} 1 & 1 & 3 \\ 0 & 2 & -1 \\ 2 & 3 & 3 \end{pmatrix}.$$

Let us carry the following row operations on the extended array

$$\begin{pmatrix} 1 & 1 & 3 & 1 & 0 & 0 \\ 0 & 2 & -1 & 0 & 1 & 0 \\ 2 & 3 & 3 & 0 & 0 & 1 \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & 3 & 1 & 0 & 0 \\ 0 & 2 & -1 & 0 & 1 & 0 \\ 0 & 1 & -3 & -2 & 0 & 1 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & 1 & 3 & 1 & 0 & 0 \\ 0 & 1 & -3 & -2 & 0 & 1 \\ 0 & 2 & -1 & 0 & 1 & 0 \end{pmatrix} \sim \begin{pmatrix} 1 & 1 & 3 & 1 & 0 & 0 \\ 0 & 1 & -3 & -2 & 0 & 1 \\ 0 & 0 & 5 & 4 & 1 & -2 \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & 1 & 3 & 1 & 0 & 0 \\ 0 & 1 & -3 & -2 & 0 & 1 \\ 0 & 0 & 1 & \frac{4}{5} & \frac{1}{5} & \frac{-2}{5} \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 6 & 3 & 0 & -1 \\ 0 & 1 & -3 & -2 & 0 & 1 \\ 0 & 0 & 1 & \frac{4}{5} & \frac{1}{5} & \frac{-2}{5} \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & 0 & 0 & 3-\frac{24}{5} & -\frac{6}{5} & -1+\frac{12}{5} \\ 0 & 1 & 0 & -2+\frac{12}{5} & \frac{3}{5} & 1-\frac{6}{5} \\ 0 & 0 & 1 & \frac{4}{5} & \frac{1}{5} & -\frac{2}{5} \end{pmatrix}.$$

Hence

$$A^{-1} = \frac{1}{5}\begin{pmatrix} -9 & -6 & -7 \\ 2 & 3 & -1 \\ 4 & 1 & -2 \end{pmatrix}.$$

3.2.3 LU Factorizations

In this subsection A will denote a square, invertible matrix, say of size $n \times n$. If $A = LU$ is the product of two triangular matrices, the first one L lower-triangular, the second one U upper-triangular, then the system $A\mathbf{x} = \mathbf{b}$ can be solved in two easy steps. Putting $U\mathbf{x} = \mathbf{y}$, we have to solve

$$A\mathbf{x} = L\,\underbrace{U\mathbf{x}}_{=\mathbf{y}} = \mathbf{b}.$$

Hence we start by solving $L\mathbf{y} = \mathbf{b}$, which is a lower-triangular system. Having done that, there remains to solve $U\mathbf{x} = \mathbf{y}$, which is an upper-triangular system.

Example. Let us solve the following system

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 3 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}.$$

We first solve

$$\begin{pmatrix}1&0&0\\1&1&0\\1&1&1\end{pmatrix}\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix}=\begin{pmatrix}b_1\\b_2\\b_3\end{pmatrix}.$$

This system is simply

$$\begin{cases}y_1 &= b_1\\ y_1+y_2 &= b_2\\ y_1+y_2+y_3 &= b_3.\end{cases}$$

We see successively (working from top to bottom)

$$\begin{cases}y_1=b_1\\ y_2=b_2-y_1=-b_1+b_2\\ y_3=b_3-y_1-y_2=-b_2+b_3.\end{cases}$$

Let us check this: Indeed

$$\begin{pmatrix}1&0&0\\1&1&0\\1&1&1\end{pmatrix}\begin{pmatrix}b_1\\-b_1+b_2\\-b_2+b_3\end{pmatrix}=\begin{pmatrix}b_1\\b_2\\b_3\end{pmatrix}.$$

Then we have to solve the system

$$\begin{pmatrix}1&1&1\\0&1&1\\0&0&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}y_1\\y_2\\y_3\end{pmatrix},$$

namely, with the found values of the y_i's

$$\begin{pmatrix}1&1&1\\0&1&1\\0&0&1\end{pmatrix}\begin{pmatrix}x_1\\x_2\\x_3\end{pmatrix}=\begin{pmatrix}b_1\\-b_1+b_2\\-b_2+b_3\end{pmatrix}.$$

This is the triangular system

$$\begin{cases}x_1+x_2+x_3 &= b_1\\ \phantom{x_1+{}}x_2+x_3 &= -b_1+b_2\\ \phantom{x_1+x_2+{}}x_3 &= -b_2+b_3.\end{cases}$$

As we are used to, we now work by back-substitution (from bottom to top)

$$\begin{cases}x_3=-b_2+b_3\\ x_2=-b_1+b_2-(-b_2+b_3)\\ \quad\;=-b_1+2b_2-b_3\\ x_1=b_1-(-b_1+2b_2-b_3)-(-b_2+b_3)\\ \quad\;=2b_1-b_2.\end{cases}$$

We may now verify

$$\begin{pmatrix}1&1&1\\1&2&2\\1&2&3\end{pmatrix}\begin{pmatrix}2b_1-b_2\\-b_1+2b_2-b_3\\-b_2+b_3\end{pmatrix}=\begin{pmatrix}b_1\\b_2\\b_3\end{pmatrix}.$$

The conclusion is the following

If we know a decomposition of A as a product LU (L lower-triangular matrix, U upper-triangular matrix), then the system $A\mathbf{x} = \mathbf{b}$ is equivalent to two successive triangular systems.

The next question is then: How can one find the factors L and U from A (when this is possible)? Here is an answer.

If it is possible to bring A into row-reduced form U, only using operations consisting of "adding to a row a multiple of a previous one," then $A = LU$ with U equal to this row-reduced form, and L is upper-triangular with 1's on its main diagonal.

This is easy to see. Indeed, the row operations in question involve multiplications by matrices $E_{ij}(c)$ which are of the mentioned form, and so is their product E:

$$A \sim EA = U \quad \text{implies} \quad A = E^{-1}U = L_1 U.$$

We can take

$$L_1 = E^{-1} : \quad \text{product in the reverse order of the } E_{ij}(-c),$$

also lower-triangular, with ones on its main diagonal. More explicitly, if the ith row $\rho_i(U)$ of U is obtained from ρ_i (ith row of A) by successive subtractions of previous pivot rows

$$\rho_i(U) = \rho_i - \ell_{i1}(\text{1st pivot row}) - \ell_{i2}(\text{2nd pivot row}) - \cdots,$$

then we have conversely

$$\rho_i = \rho_i(U) + \ell_{i1}(\text{1st pivot row}) + \ell_{i2}(\text{2nd pivot row}) + \cdots,$$

namely since the pivot rows are also the rows of U (when a pivot row is reached, it is kept constant in the procedure)

$$\begin{aligned} \rho_i &= \rho_i(U) + \ell_{i1}\rho_1(U) + \ell_{i2}\rho_2(U) + \cdots \\ &= \sum_{k<i} \ell_{ik}\rho_k(U) + \rho_i(U). \end{aligned}$$

This is exactly the ith row of LU where the coefficients of the ith row of L are the multipliers used in the elimination: The ith coefficient of L is 1, and $\ell_{ij} = 0$ for $j > i$. This procedure calls for several comments.

(1) Recall that the row operations used to bring A in a row-reduced form are not uniquely determined. This is reflected in the result in the following way. It is easy to factor $U = DU_1$ where D is diagonal (contains the pivot entries), and U_1 is upper-triangular with 1's on its main diagonal. Then $A = LU_1$ with $L = L_1 D$ is lower-triangular. (More generally, one can transfer a partial number of pivot entries from U to L.) It would be better to write

$$A = L_1 D U_1,$$

with both L_1 and U_1 having 1's on their main diagonal, while D is the diagonal matrix formed with the pivots entries p_i's. The three factors are then uniquely determined (when they exist).
(2) It may be impossible to bring A into row-reduced form without exchange of rows. For example, this is the case for the matrix

$$\begin{pmatrix} 0 & 1 \\ 1 & d \end{pmatrix}.$$

But using row operations of the type mentioned, we may always attain a form

$$V = \begin{pmatrix} & & & \boxed{q_1} & \\ 0 & & & \cdots & 0 \\ \boxed{q_2} & & & & \\ & & \boxed{q_3} & & \\ & & & 0\cdots & 0 \\ & \boxed{q_r} & & & \\ & 0 & & & 0 \end{pmatrix},$$

where all entries *below* a pivot $q_i (\neq 0)$ are zero. A permutation matrix P (product of elementary P_{ij}'s, corresponding to exchanges of rows) will now bring this matrix in row-reduced form

$$V \sim PV = U.$$

Reversing the row operations, we find

$$\begin{aligned} A &\sim EA = V \sim PV = U, \\ A &= E^{-1}V = E^{-1}P^{-1}U = LWU, \end{aligned}$$

where $W = P^{-1}$ is a permutation matrix (having the same rows as the identity matrix, in a possibly different order). Let us show how to proceed to reach V on an example:

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 2 & 4 & 3 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix} \sim \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & -1 & 2 & 5 \end{pmatrix} \sim \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & -1 & 0 & 3 \end{pmatrix}$$

$$\sim \begin{pmatrix} 0 & 0 & 0 & 0 & \boxed{1} \\ 0 & \boxed{1} & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \boxed{1} & 0 \\ \boxed{1} & 0 & -1 & 0 & 0 \end{pmatrix}$$

(the q_i's are the framed coefficients).

The final question is: How can we recognize if a given invertible matrix A can be written in the LU form? An answer to this question will be given in Sec. 11.3.3.

3.2.4 Simultaneous Resolution of Linear Systems

If two linear systems having the same matrix of coefficients A are to be solved, say

$$(S) \quad A\mathbf{x} = \mathbf{b}, \qquad (S)' \quad A\mathbf{x} = \mathbf{b}',$$

we might save time by carrying out row operations on the extended matrix $(A \mid \mathbf{b}\ \mathbf{b}')$. In the same vein, we might be interested in solving the particular systems

$$(S_1) \quad A\mathbf{x} = \mathbf{e}_1, \quad \ldots, \quad (S_n) \quad A\mathbf{x} = \mathbf{e}_n,$$

where the $\mathbf{e}_i$ are the basic n-tuples. For this purpose, we may group all systems in one single extended array

$$(A \mid \mathbf{e}_1 \ldots \mathbf{e}_n).$$

The complete resolution is attained when the first part is in reduced echelon form. Assuming that this system has as many equations as variables, and has maximal rank: $r = m = n$, we find the unique solution $\mathbf{x}_j$ of (S_j) in the $(n+j)$th column of this reduced echelon form. The solution of the general system $A\mathbf{x} = \mathbf{b} = (b_i)$ is the linear combination

$$b_1\mathbf{x}_1 + \cdots + b_n\mathbf{x}_n$$

of these particular solutions. With matrix multiplication, a regular system $A\mathbf{x} = \mathbf{b}$ is solved by $\mathbf{x} = A^{-1}\mathbf{b}$. This row by column multiplication shows that the columns of A^{-1} are precisely the particular solutions $\mathbf{x}_i$.

3.3 Matrix Multiplication by Blocks

3.3.1 Explanation of the Method

In a matrix product AB, let us consider B as an ntuple of rows ρ_i. Here is a particular case showing how the product is computed:

$$(a\ b\ c)\begin{pmatrix}\rho_1\\ \rho_2\\ \rho_3\end{pmatrix} = (a\ b\ c)\begin{pmatrix}\leftarrow \cdots a_{1j} \cdots \rightarrow\\ \leftarrow \cdots a_{2j} \cdots \rightarrow\\ \leftarrow \cdots a_{3j} \cdots \rightarrow\end{pmatrix} = a\rho_1 + b\rho_2 + c\rho_3,$$

just as if the entries ρ_i were scalars. From this observation follows that

$$\begin{pmatrix}a_1 & b_1 & c_1\\ a_2 & b_2 & c_2\\ \vdots & \vdots & \vdots\end{pmatrix}\begin{pmatrix}\rho_1\\ \rho_2\\ \rho_3\end{pmatrix} = \begin{pmatrix}a_1\rho_1 + b_1\rho_2 + c_1\rho_3\\ a_2\rho_1 + b_2\rho_2 + c_2\rho_3\\ \vdots\end{pmatrix}.$$

This is precisely what was systematically used with row operations. More generally, we see that

The rows of a matrix product AB are linear combinations of the rows of the second factor B.

We used this for the interpretation of row operations by left matrix multiplication $A \sim U = EA$. Treating rows formally as scalars in a matrix product, is a first case of *block multiplication*. We intend to show how this principle works quite generally.

By definition, the matrix product of two matrices A and B of adapted sizes, is given by

$$A(\mathbf{b}_1 \cdots \mathbf{b}_n) = (A\mathbf{b}_1 \cdots A\mathbf{b}_n),$$

where $\mathbf{b}_j$ is the jth column of B. Let us consider explicitly the simplest case, namely $B = \mathbf{x}$ has only one column. Instead of considering A as $n$tuple of rows

$$A\mathbf{x} = \begin{pmatrix} a_{11}x_1 & + & \cdots & + & a_{1n}x_n \\ a_{21}x_1 & + & \cdots & + & a_{2n}x_n \\ \vdots & & & & \vdots \\ a_{m1}x_1 & + & \cdots & + & a_{mn}x_n \end{pmatrix} = \begin{pmatrix} \rho_1\mathbf{x} \\ \rho_2\mathbf{x} \\ \vdots \\ \rho_n\mathbf{x} \end{pmatrix} = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_n \end{pmatrix} \mathbf{x},$$

introduce its columns $\mathbf{a}_j$, so that

$$A\mathbf{x} = \begin{pmatrix} a_{11}x_1 & + & \cdots & + & a_{1n}x_n \\ a_{21}x_1 & + & \cdots & + & a_{2n}x_n \\ \vdots & & & & \vdots \\ a_{m1}x_1 & + & \cdots & + & a_{mn}x_n \end{pmatrix} = \sum_j x_j\mathbf{a}_j = (\mathbf{a}_1 \cdots \mathbf{a}_n)\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix},$$

formally as if the columns of A were scalars. The same method works when B has several columns

$$(\mathbf{a}_1 \cdots \mathbf{a}_n)\begin{pmatrix} x_1 & y_1 & z_1 & \cdots \\ \vdots & \vdots & \vdots & \\ x_n & y_n & z_n & \cdots \end{pmatrix} = \left(\sum_j x_j\mathbf{a}_j \;\; \sum_j y_j\mathbf{a}_j \;\; \sum_j z_j\mathbf{a}_j \;\; \cdots \right)$$

We infer the following general principle

The columns of a matrix product AB are linear combinations of the columns of the first factor A.

The possibility of the block multiplication

$$A(B_1 \mid B_2 \mid \ldots \mid B_n) = (AB_1 \mid AB_2 \mid \ldots \mid AB_n)$$

follows from the definition of the matrix product. This was used for the inversion algorithm, when we wrote

$$\begin{aligned} (A \mid I) &\sim (U \mid *) = E(A \mid I) = (EA \mid E) \\ &\sim F(EA \mid E) = (FEA \mid FE) = (I \mid FE) = (I \mid A^{-1}). \end{aligned}$$

Symmetrically, it is obvious that a block multiplication is valid when the first matrix is partitioned as follows

$$\begin{pmatrix} A_1 \\ \cdots \\ A_n \end{pmatrix} B = \begin{pmatrix} A_1 B \\ \cdots \\ A_n B \end{pmatrix}.$$

More generally, we see that

$$\begin{pmatrix} A_1 \\ \cdots \\ A_n \end{pmatrix} (B_1 \mid B_2 \mid B_3) = \begin{pmatrix} A_1 B_1 & A_1 B_2 & A_1 B_3 \\ \cdots & \cdots & \cdots \\ A_n B_1 & A_n B_2 & A_n B_3 \end{pmatrix}.$$

Coming back to

$$\begin{pmatrix} A_1 \\ A_2 \end{pmatrix} B = \begin{pmatrix} A_1 B \\ A_2 B \end{pmatrix},$$

we may introduce a partition of A into p and q columns, with a corresponding partition of B into p and q rows. We then have

$$\begin{pmatrix} A_1' \mid A_1'' \\ A_2' \mid A_2'' \end{pmatrix} \begin{pmatrix} B' \\ B'' \end{pmatrix} = \begin{pmatrix} A_1' B' + A_1'' B'' \\ A_2' B' + A_2'' B'' \end{pmatrix},$$

simply since the row by column multiplications may be decomposed as

$$\sum_{1 \leqslant k \leqslant n} a_{ik} b_{kj} = \sum_{k \leqslant p} a_{ik} b_{kj} + \sum_{k > p} a_{ik} b_{kj}.$$

If the B matrices are also decomposed by a vertical separation, we conclude

$$\begin{pmatrix} A_1' \mid A_1'' \\ A_2' \mid A_2'' \end{pmatrix} \begin{pmatrix} B_1' \mid B_2' \\ B_1'' \mid B_2'' \end{pmatrix} = \begin{pmatrix} A_1' B_1' + A_1'' B_1'' \mid A_1' B_2' + A_1'' B_2'' \\ A_2' B_1' + A_2'' B_1'' \mid A_2' B_2' + A_2'' B_2'' \end{pmatrix}.$$

This is exactly the same formula as for the product of 2×2 matrices: Formally, the blocks are treated as scalars, but the order in which they are multiplied is important. Fast multiplication algorithms are based on this possibility (see 3.4.3). For example, if $A_2' = 0$ and $B_1'' = 0$, we get

$$\begin{pmatrix} A_1' & A_1'' \\ 0 & A_2'' \end{pmatrix} \begin{pmatrix} B_1' & B_2' \\ 0 & B_2'' \end{pmatrix} = \begin{pmatrix} A_1' B_1' & A_1' B_2' + A_1'' B_2'' \\ 0 & A_2'' B_2'' \end{pmatrix}.$$

More general cases with an arbitrary number of blocks (of suitable sizes) are obviously valid. Let us simply summarize these observations in the following way:

> ***Row by column multiplication can be made when the matrices are decomposed into blocks of compatible sizes.***

3.3.2 The Field of Complex Numbers

Let us solve the equation $X^2 = -I$ in the algebra of 2×2 real matrices: We are looking for the matrices $X = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ such that

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a^2+bc & ab+bd \\ ac+cd & bc+d^2 \end{pmatrix} = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Here are the conditions

$$\begin{cases} a^2+bc = -1, & (a+d)b = 0, \\ (a+d)c = 0, & bc+d^2 = -1. \end{cases}$$

The first equation shows that $bc = -1 - a^2 \leqslant -1 < 0$ hence b and c are nonzero and have opposite signs (the same conclusion follows from the last equation). The next two equations furnish $a + d = 0$, or $d = -a$ (hence $a^2 = d^2$). If we take arbitrary values for a and $b \neq 0$, and then

$$c = -(1+a^2)/b \text{ and } d = -a,$$

we find infinitely many matrices

$$X = \begin{pmatrix} a & b \\ -\frac{1+a^2}{b} & -a \end{pmatrix} \quad (b \neq 0),$$

which are all solutions of $X^2 = -I$. (In a *noncommutative* algebra, a quadratic equation may have infinitely many solutions.) Let us choose one of them, say

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

Then we can consider the matrices of the form

$$\begin{pmatrix} a & b \\ -b & a \end{pmatrix} = aI + bJ.$$

They form a subspace of the space of 2×2 matrices. Since I, J is a basis of this subspace, it has dimension 2. The product of two matrices $aI + bJ$ and $\alpha I + \beta J$ again has the same form:

$$\begin{aligned} (aI+bJ)(\alpha I+\beta J) &= a\alpha I + (a\beta + b\alpha)J + b\beta J^2 \\ &= (a\alpha - b\beta)I + (a\beta + b\alpha)J. \end{aligned}$$

Observe that

$$(aI+bJ)(aI-bJ) = (a^2+b^2)I$$

is nonzero when $aI + bJ \neq 0$. Hence $aI + bJ \neq 0$ implies

$$aI + bJ \text{ invertible and } (aI+bJ)^{-1} = \frac{a}{a^2+b^2}I + \frac{-b}{a^2+b^2}J.$$

We obtain thus a model for the *complex numbers*. Instead of $aI + bJ$, we write $z = a + ib$, and define

$$\begin{array}{rcll} a & = & \Re(z) & \textit{real part of } z, \\ b & = & \Im(z) & \textit{imaginary part of } z, \\ \overline{z} & = & a - bi & (\textit{complex}) \textit{ conjugate of } z, \\ |z| & = & \sqrt{a^2+b^2} & \textit{absolute value of } z. \end{array}$$

As we have seen, $|z| = 0$ only when $z = 0$ ($a = 0$ and $b = 0$), and any $z \neq 0$ is invertible with

$$z^{-1} = \frac{1}{z} = \frac{\overline{z}}{|z|^2} = \frac{\Re(z)}{|z|^2} - i\frac{\Im(z)}{|z|^2}.$$

The set of complex numbers is the *complex field* $\mathbf{C}$. We view it as a plane, identifying the basis I, J to the canonical basis of $\mathbf{R}^2$.

3.4 Appendix

3.4.1 Affine Maps

A matrix A of size $m \times n$ defines a map

$$\begin{array}{ccc} \mathbf{R}^n & \longrightarrow & \mathbf{R}^m \\ \mathbf{x} & \longmapsto & A\mathbf{x}, \end{array}$$

also denoted by A. Such a map has the basic property $A\mathbf{0} = \mathbf{0}$. Any map $f : \mathbf{R}^n \to \mathbf{R}^m$ of the form

$$\mathbf{x} \longmapsto f(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$$

for some $\mathbf{b} \in \mathbf{R}^m$, is called an *affine map*. Since $f(\mathbf{0}) = \mathbf{b}$, such a map can be given by a matrix multiplication only when $\mathbf{b} = \mathbf{0}$. However, one may also view affine maps as being induced by matrix multiplication as follows. The matrices of size $(m+1) \times (n+1)$

$$\left(\begin{array}{c|c} A & \mathbf{b} \\ \hline 0 \dots 0 & 1 \end{array}\right)$$

define maps $\mathbf{R}^{n+1} \to \mathbf{R}^{m+1}$. On the subset V_1 consisting of vectors with last component 1, we recover the affine maps

$$\left(\begin{array}{c|c} A & \mathbf{b} \\ \hline 0 \dots 0 & 1 \end{array}\right) \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} = \begin{pmatrix} A\mathbf{x} + \mathbf{b} \\ 1 \end{pmatrix}.$$

Since V_1 does not contain the 0 vector, it is not a vector subspace. But if V denotes the *subspace* consisting of vectors having last component 0, then

$$V_1 = \{\mathbf{t} + \mathbf{v} : \mathbf{v} \in V\} = \mathbf{t} + V,$$

where $\mathbf{t}$ denotes any vector having last component 1. We view it as a *translate* of a vector subspace. Any subset of a vector space which is obtained by translation from a vector subspace is called *affine subspace*. For example, the set of solutions of a linear system is an affine subspace: It is a translate of the subspace of solutions of the associated homogeneous system.

Notice how affine maps compose

$$\left(\begin{array}{c|c} A & \mathbf{b} \\ \hline 0 \ldots 0 & 1 \end{array}\right)\left(\begin{array}{c|c} A' & \mathbf{b}' \\ \hline 0 \ldots 0 & 1 \end{array}\right) = \left(\begin{array}{c|c} AB & A\mathbf{b}' + \mathbf{b} \\ \hline 0 \ \ldots \ 0 & 1 \end{array}\right).$$

3.4.2 The Field of Quaternions

It is possible to construct a larger field than $\mathbf{C}$, namely the *field of quaternions* $\mathbf{H}$. However, in this field ab may differ from ba: This field is *not commutative* (hence is not a good candidate for a field of scalars of vector spaces). Nevertheless, its construction may be based on block multiplication of matrices, and therefore furnishes an interesting example of this method. Let us start with the four basic matrices

$$\mathbf{1} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}, \quad \mathbf{i} = \begin{pmatrix} 0 & J \\ J & 0 \end{pmatrix}, \quad \mathbf{j} = \begin{pmatrix} 0 & I \\ -I & 0 \end{pmatrix}, \quad \mathbf{k} = \mathbf{i}\,\mathbf{j} = \begin{pmatrix} -J & 0 \\ 0 & J \end{pmatrix}.$$

They satisfy

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -I, \quad \mathbf{i}\,\mathbf{j} = \mathbf{k} = -\mathbf{j}\,\mathbf{i} \quad \text{and cyclic permutations.}$$

Embedding 2×2 complex matrices by blocks in 4×4 real matrices, we get injective maps

$$\mathbf{H} \longrightarrow M_2(\mathbf{C}) \longrightarrow M_4(\mathbf{R}).$$

A quaternion is an expression

$$\mathbf{q} = a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} \quad (a, b, c, d \in \mathbf{R}).$$

As with complex numbers, a conjugation can be introduced for quaternions

$$\overline{\mathbf{q}} = a\mathbf{1} - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}.$$

One can check

$$\overline{\mathbf{q}}\,\mathbf{q} = \mathbf{q}\,\overline{\mathbf{q}} = a^2 + b^2 + c^2 + d^2.$$

From this follows that any nonzero quaternion $\mathbf{q}$ is invertible with

$$\mathbf{q}^{-1} = \overline{\mathbf{q}}/N \quad (N = \overline{\mathbf{q}}\,\mathbf{q} \neq 0).$$

Hence $\mathbf{H}$ is a *noncommutative field.*

3.4.3 The Strassen Algorithm

The product of two 2×2 matrices is given by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} = \begin{pmatrix} a\alpha + b\gamma & a\beta + b\delta \\ c\alpha + d\gamma & c\beta + d\delta \end{pmatrix} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

We find

$$A = p + s + t - v, \ B = r + v, \ C = q + t, \ D = p - q + r + u,$$

where

$$\begin{cases} p &= (a+d)(\alpha+\delta) \\ q &= (c+d)\alpha \\ r &= a(\beta-\delta) \\ s &= (b-d)(\gamma+\delta) \\ t &= d(\gamma-\alpha) \\ u &= (c-a)(\beta+\alpha) \\ v &= (b+a)\delta. \end{cases}$$

With this method, only 7 multiplications—instead of 8—are needed (it is true that 18 sums occur, but with some fast processors, the time to compute them may be neglected). This tiny advantage becomes more and more important if it is used repeatedly: A square matrix of size 1024×1024 can be considered as a 2×2 matrix of blocks, each of size 512×512, etc. Observe that the above formulas are valid even for a noncommutative product, which is the case with blocks. For general $n \times n$ matrices, where n is not a power of 2, it is advantageous to complete the matrix by 0's up to the nearest $2^m \times 2^m$ size. Hence the method is useful for large size matrices.

Here is an estimate of the time cost for this multiplication method (neglecting time taken by additions). Let us only consider the case where $n = 2^m$. Then, instead of $n^3 = 2^{3m} = 8^m$ operations, only

$$7^m = 2^{m \log_2 7} = n^{\log_2 7}$$

will be required. The exponent is the logarithm to the base 2 of 7

$$\log_2 7 = \frac{\log 7}{\log 2} = 2.80735\ldots$$

The Strassen algorithm has *order* $\mathcal{O}(n^{2.8\ldots})$.

3.5 Exercises

1. Let $\vec{a}, \vec{b} \in \mathbf{R}^3$ be linearly independent. Find all vectors $\vec{x}, \vec{y} \in \mathbf{R}^3$ such that

$$\begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{pmatrix} \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \\ a_3 & b_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

2. Draw the image of the unit square by the linear map having matrix

$$\begin{pmatrix} \cos\varphi & \sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}$$

in the canonical basis of $\mathbf{R}^2$. Compute M^2, and the nth power of M (to simplify the notation, you may put $\cos\varphi = a$ and $\sin\varphi = b$).

3. Compute the square and the cube of

$$A = \begin{pmatrix} a & b \\ 0 & a \end{pmatrix}.$$

Can you guess the form of A^n? Prove your guess by induction.

4. Solve the equation $X^2 = I$ in the algebra $\mathbf{M}_2(\mathbf{R})$ consisting of 2×2 real matrices.

5. Let

$$a(t) = \begin{pmatrix} t & 0 \\ 0 & 1/t \end{pmatrix}, \quad b(s) = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix}.$$

Compute $a(t)b(s)a(t)^{-1}$. With

$$\alpha(t) = \begin{pmatrix} t & 0 \\ 0 & -t \end{pmatrix}, \quad \beta(s) = \begin{pmatrix} 0 & s \\ 0 & 0 \end{pmatrix},$$

compute $\alpha(t)\beta(s) - \beta(s)\alpha(t)$.

6. Show that any matrix of size $2n \times 2n$ of the form $T = \begin{pmatrix} 0 & X \\ 0 & 0 \end{pmatrix}$ (where X is any $n \times n$ matrix) satisfies $T^2 = 0$. When $T^2 = 0$, prove that the inverse of $I + T$ is $I - T$.

7. Compute all powers of the matrices

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Any square matrix T (or linear map $T : E \to E$) such that $T^k = 0$ for some positive integer k, is called *nilpotent*. If $T^k = 0$, show that $I + T + \cdots + T^{k-1}$ is the inverse of $I - T$.

8. Compute all powers of the matrix

$$\begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

9. Are the following matrices row-reduced? triangular?

$$\begin{pmatrix} 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Prove carefully the implications (for square matrices)

$$\text{row-reduced} \implies \text{upper-triangular},$$

$$\text{upper-triangular and maximal rank} \implies \text{row-reduced}.$$

10. Using suitable row operations, determine the inverse of the matrices

$$\begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ b & c & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & 5 & 6 \\ 0 & 0 & 1 & 7 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

11. Compute the first twelve powers (!) of

$$A = \begin{pmatrix} 2 & -16 & 3 & -1 \\ 1 & -2 & 0 & 0 \\ 4 & 5 & -3 & 1 \\ 0 & 35 & -8 & 3 \end{pmatrix}.$$

12. The inverse of $\begin{pmatrix} 0 & 1 \\ 1 & a \end{pmatrix}$ is $\begin{pmatrix} -a & 1 \\ 1 & 0 \end{pmatrix}$. More generally, find the inverse of

$$\begin{pmatrix} 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ 1 & a_1 & \dots & a_n \end{pmatrix} = \begin{pmatrix} 0 & \boxed{\quad I_n \quad} \\ 1 & {}^t\vec{a} \end{pmatrix} \qquad (\text{size } (n+1)\times(n+1)\ ).$$

13. What is the inverse of a square matrix

$$\begin{pmatrix} 0 & 0 & \dots & 0 & a_n \\ 0 & & \dots & a_{n-1} & 0 \\ \vdots & & \iddots & & \vdots \\ 0 & a_2 & \dots & & 0 \\ a_1 & 0 & \dots & 0 & 0 \end{pmatrix}.$$

14. Is there a 2×2 matrix X such that $X^2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$? Is there a 3×3 matrix X such that

$$X^2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}?$$

15. Let A be any square matrix. Prove

$$\ker A \subset \ker A^2 \subset \ker A^3 \subset \cdots$$

If $\ker A = \ker A^2$, prove that all subsequent inclusions are equalities:

$$\ker A^n = \ker A^{n+1} \quad (n \geqslant 1).$$

Use the preceding observation to prove that there is no 3×3 matrix X such that

$$X^2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

16. Let x_i denote three distinct real numbers. Show that for each choice of distinct values y_i, there is one and only one linear fractional transformation f such that $f(x_i) = y_i$ $(i = 1, 2, 3)$.

17. Let $B = (b_{ij})_{i,j \geqslant 0}$ be the lower triangular matrix containing the binomial coefficients $b_{ij} = \binom{i}{j}$

$$B = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots \\ 1 & 1 & 0 & 0 & \dots \\ 1 & 2 & 1 & 0 & \dots \\ 1 & 3 & 3 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Show that $B^2 = DBD^{-1}$ where D is the diagonal matrix having entries 1, 2, 4,...,2^i,... in its diagonal. (We may consider that B has infinite size, since its triangular form furnishes each coefficient of B^2 as a finite sum; the corresponding result contains its analogues for all finite $N \times N$ sizes.) Show that this implies

$$B^{2^n} = D^n B D^{-n} \quad (n \geqslant 0).$$

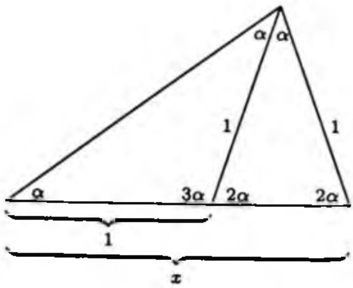
Hint: Use the easily proved relations

$$\begin{aligned} \binom{i}{j}\binom{j}{k} &= \binom{i}{k}\binom{i-k}{i-j}, \\ \sum_{0 \leqslant \ell \leqslant i-k} \binom{i-k}{\ell} &= 2^{i-k}. \end{aligned}$$

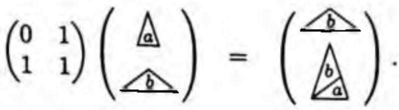
18. Compute

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}^n \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} a_n \\ b_n \end{pmatrix}$$

for $n = 1, 2, 3$, and 4. Use triangles similar to



(where $\alpha = \pi/5$, and x is determined by $\frac{x}{1} = \frac{1}{x-1}$, hence is the positive root of $x^2 = x + 1$) to illustrate the multiplications geometrically as in the following picture



Notes

The matrix multiplication for pixel matrices was (first?) systematically used by Jean-Pierre Reveillès in his PhD. The digitalization of a straight line is not obvious, particularly when its slope is irrational. The Bresenham algorithm shows how to do it: See

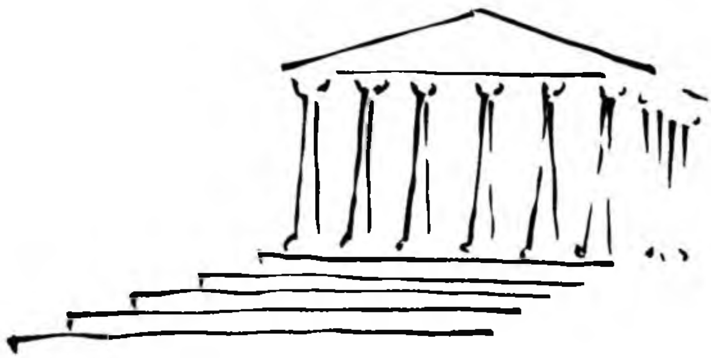
> J.-P. Reveillès: Géométrie et ordinateurs I et II: droites, cercles et paraboles, Gazette des mathématiciens de la SMF **78** (1998), 31-49, **79** (1999), 29-44.

The positive root of $x^2 = x + 1$ is the *golden ratio*. Its approximate value is $x = 1.6180\ldots$. In Chapter 6, we shall see its relation to the Fibonacci numbers.

For the matrix multiplication in connection with games and puzzles (exercise 18), see www.trigam.ch by J. Bauer and J.-P. Lebet.

Quest for an identity:

$$\textit{To be} \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}, \quad \textit{or not to be} \begin{pmatrix} 0 & & 1 \\ & \iddots & \\ 1 & & 0 \end{pmatrix}!$$

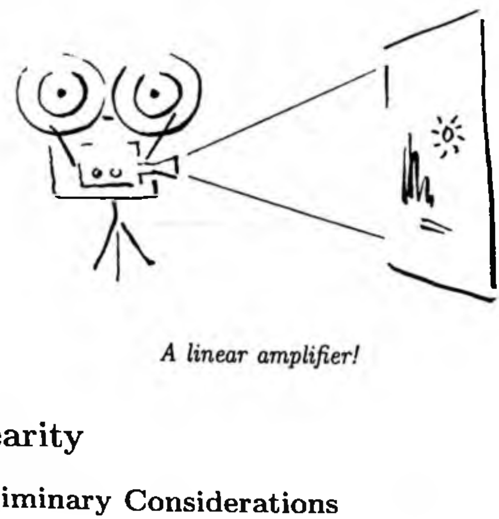


Remember: First rows, then columns!

Chapter 4

Linear Maps

The concept of linearity has made its way in everyday life. One speaks of linear growth, linear amplifier, etc. Linearity first refers to proportionality between causes and consequences.



A linear amplifier!

4.1 Linearity

4.1.1 Preliminary Considerations

(1) Here is a familiar example: A bank delivers a fixed interest rate on deposits. After a certain time lapse, an initial capital C yields an interest $I = f(C)$. We represent this situation by a box—the bank—acting as a function, producing the interest:

$$C \to \boxed{\text{CITYBANK}} \to I.$$

Proportional deposits will produce proportional interests. Moreover, if another capital C' yields the interest I' under the same circumstances, then the capital $C + C'$ will produce the interest $I + I'$ on the same account. These facts are used in the *false assumption method* for finding which capital C placed during three years at a rate 5%, bears an interest of say \$150. Here we *assume* that an initial capital of \$100 is placed under the same conditions and find an interest of \$15. Then we conclude that if we have obtained ten times this sum, it is because the initial capital was ten times larger: $C = \$1000$.

Observe however that if the data

$$C\text{: capital, } I_r\text{: interest rate, } T\text{: time}$$

are grouped in a vector, then we obtain a nonlinear box

$$\begin{pmatrix} C \\ I_r \\ T \end{pmatrix} \to \boxed{\text{CITYBANK}} \to I.$$

Indeed, when C, I_r, and T are simultaneously multiplied by a factor a, the interest yielded is multiplied by a factor a^3, instead of the factor a as linearity would require.

(2) Many physical laws have a linear character:

- The intensity I of an electrical current in a conducting medium is proportional to the applied tension V (voltage)

$$V = RI \quad \text{(Ohm's law)}.$$

  Here, R is the resistance of the considered medium.

- For one mole of a perfect gas in a fixed volume V, the pressure p is proportional to the (absolute) temperature T

$$pV = RT \quad \text{(perfect gases law)}.$$

  Here $R = Nk$ (N: Avogadro's number, k: Boltzmann's constant).

- The deformation $\Delta\ell$ of a "solid" is proportional to the force acting on it. More precisely, the relative elongation in percent $\Delta\ell/\ell$ of a certain rod of length ℓ is proportional to the tension (constraint) σ applied to it:

$$\Delta\ell/\ell = \sigma/E \quad \text{(Hooke's law)}.$$

(3) In control theory, certain physical systems called plants, are studied. A plant is a device having an output depending on an input. Typically, the plant could be a thermostat controlling the temperature in a car, where the input would consist of both outside and inside temperature, and output is directed to the ventilation and climate control system of the car. Another example of plant is a linear amplifier, namely an electronic device which amplifies a signal in a faithful way, at least in a certain domain of frequencies. Such a plant can be considered as a *black box*, having a certain number of input ports, where the signals s_i enter, and a (possibly different) number of exit ports producing outputs e_j. The plant has *linear characteristics* when:

- ➢ An amplified signal produces an amplified output by the same factor
- ➢ Superposed input signals produce superposed outputs.

If the action of the black box is represented by a map f

$$\mathbf{s} = (s_i) \to \boxed{f} \to \mathbf{e} = (e_j),$$

its linearity characteristics may be formalized as follows:

$$f(a\mathbf{s}) = a\mathbf{e} = af(\mathbf{s}) \quad (\textit{homogeneity})$$
$$f(\mathbf{s}_1 + \mathbf{s}_2) = \mathbf{e}_1 + \mathbf{e}_2 = f(\mathbf{s}_1) + f(\mathbf{s}_2) \quad (\textit{additivity}).$$

4.1.2 Definition and First Properties

Let us now come to a precise mathematical definition.

Definition. *A mapping $f : E \to F$ between two vector spaces is* linear *when the following properties hold*

$$f(a\mathbf{x}) = af(\mathbf{x}), \quad f(\mathbf{x}+\mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y}) \quad (\mathbf{x}, \mathbf{y} \in E,\ a \text{ scalar}).$$

Taking $a = 0$ in $f(a\mathbf{x}) = af(\mathbf{x})$, we see that it implies $f(0_E) = 0_F$ (0_E stands for the zero element of E, 0_F for the zero element of F). Instead of $0_E \in E, \dots$ we shall simply write 0, relying on the reader for a proper understanding of which zero element appears. Also observe that the two required identities can be gathered into a single one

$$f(a\mathbf{x} + \mathbf{y}) = af(\mathbf{x}) + f(\mathbf{y}).$$

Indeed, $\mathbf{y} = 0$ gives homogeneity, while $a = 1$ gives additivity. Of course, this relation implies

$$f(a\mathbf{x} + b\mathbf{y}) = af(\mathbf{x}) + bf(\mathbf{y}),$$

and more generally

$$f\Big(\sum_{1\leqslant i\leqslant n} a_i\mathbf{x}_i\Big) = \sum_{1\leqslant i\leqslant n} a_i f(\mathbf{x}_i)$$

(as an induction on the number n of terms in the sum shows), hence:

> *Linear maps transform linear combinations into linear combinations having the same coefficients.*

Geometrical Interpretation. The characterization of linearity by $f(a\mathbf{x}+\mathbf{y}) = af(\mathbf{x}) + f(\mathbf{y})$ has the following geometrical meaning. Interpret the scalar a as "time" t and consider $\mathbf{v}_t = t\mathbf{x} + \mathbf{y}$ as giving a parameterization of the straight line going through $\mathbf{y}$ ($t = 0$) having direction given by $\mathbf{x}$ (at least if this vector is nonzero). Then the images $f(\mathbf{v}_t)$ should also constitute a parameterization $tf(\mathbf{x}) + f(\mathbf{y})$ of a straight line based at $f(\mathbf{y})$ ($t = 0$) and having direction given by $f(\mathbf{x})$ (same comment). Hence:

Linear maps transform straight lines into straight lines (or points).

Comment. A linear map taking values in the one-dimensional vector space consisting of the scalars is usually called a *linear form*. This terminology comes from the use of *form* for any *homogeneous function* $f : E \to \mathbf{R}$ taking scalar values. We shall study *quadratic forms* on a vector space E, namely functions $f : E \to \mathbf{R}$ satisfying

$$f(a\mathbf{x}) = a^2 f(\mathbf{x}) \quad (a \text{ scalar}).$$

Similarly, *a cubic form* f on E satisfies

$$f(a\mathbf{x}) = a^3 f(\mathbf{x}) \quad (a \text{ scalar}).$$

The function $(x, y, z) \mapsto f(x, y, z) = (xyz)^{1/3}$ is homogeneous of degree 1 since

$$f(ax, ay, az) = (a^3)^{1/3}(xyz)^{1/3} = a(xyz)^{1/3} = af(x, y, z).$$

But note that $g(x, y) = \sqrt{x^2 + y^2}$ is only *positively* homogeneous:

$$g(ax, ay) = \sqrt{a^2x^2 + a^2y^2} = |a|\sqrt{x^2 + y^2} = |a|g(x, y).$$

4.1.3 Examples of Linear Maps

(1) A map $f : \mathbf{R} \to \mathbf{R}$ of the form $f(x) = ax + b$ is linear only if $b = 0$. Indeed, linearity requires $f(0) = 0$. When $b \neq 0$, the function $f(x) = ax + b$ is called *affine*, or badly enough *affine linear* (see Sec. 3.4.1) (hence an affine linear map is *not* linear in the sense of linear algebra).

(2) In Sec. 3.1.4, we have considered the linear map

$$\begin{array}{rccc} L_A : & \mathbf{R}^n & \longrightarrow & \mathbf{R}^m \\ & \mathbf{x} & \longmapsto & A\mathbf{x}, \end{array}$$

produced by left multiplication by a matrix A of size $m \times n$. The characteristic property of linear maps

$$L_A(a\mathbf{x} + \mathbf{y}) = A(a\mathbf{x} + \mathbf{y}) = aA\mathbf{x} + A\mathbf{y} = aL_A(\mathbf{x}) + L_A(\mathbf{y})$$

is satisfied (matrix multiplication is distributive). We shall prove that *any linear map from* $\mathbf{R}^n$ *to* $\mathbf{R}^m$ *is given by a matrix multiplication as before.* This example turns out to produce the most general linear map $f : \mathbf{R}^n \to \mathbf{R}^m$. Recall that the jth column of a matrix A is given by $A\mathbf{e}_j$ where $\mathbf{e}_j$ is the n-tuple having 0's except at the jth place, where it has a 1. Hence, there is only one possible matrix A for any given linear map $f : \mathbf{R}^n \to \mathbf{R}^m$. The notation $\mathbf{x} \mapsto \mathbf{y} = A\mathbf{x}$ generalizes the one dimensional case: Any linear map $\mathbf{R} \to \mathbf{R}$ is given in the form $x \mapsto y = ax$.

(3) The derivation operation is linear. This map $f \mapsto f'$ indeed has the property

$$(af + g)' = af' + g' \quad (a \text{ scalar}),$$

characteristic of linearity. We may consider this map as being defined on the vector space of polynomials, with $(x^j)' = jx^{j-1}$ ($j \geqslant 1$).

(4) The operation consisting in taking the primitive—vanishing at the origin—of continuous functions, is a linear operation. It is given by

$$f \longmapsto \int_0^x f(t)\,dt.$$

(5) The evaluation of functions at a point is linear. For example, the evaluation of polynomials at $\pi = 3.14159\ldots$ is a linear form $f \mapsto f(\pi)$. An efficient procedure for evaluating a polynomial $f(x) = a_0 + a_1 x + a_2 x^2 + \cdots$ at π is given by

$$f(\pi) = a_0 + \pi(a_1 + \pi(a_2 + \pi(\cdots)\cdots)).$$

(6) A rotation around an axis containing the origin in the usual space $\mathbf{R}^3$ is a linear operation.

4.2 General Results

Let f and g be two linear maps $E \to F$ between the same vector spaces. Then for each scalar c, the map $h = cf + g$ defined by

$$\mathbf{x} \longmapsto h(\mathbf{x}) = cf(\mathbf{x}) + g(\mathbf{x})$$

is also linear: Here is the verification of the formal identity required

$$\begin{aligned}
h(a\mathbf{x}+\mathbf{y}) &= cf(a\mathbf{x}+\mathbf{y}) + g(a\mathbf{x}+\mathbf{y}) \\
&= c(af(\mathbf{x}) + f(\mathbf{y})) \ + \ ag(\mathbf{x}) + g(\mathbf{y}) \\
&= acf(\mathbf{x}) + ag(\mathbf{x}) + cf(\mathbf{y}) + g(\mathbf{y}) \\
&= a(cf+g)(\mathbf{x}) + (cf+g)(\mathbf{y}) \\
&= ah(\mathbf{x}) + h(\mathbf{y}).
\end{aligned}$$

4.2.1 Image and Kernel of a Linear Map

For any map $f : E \to F$ between sets, and any subset $X \subset E$, we denote its image by $f(X) = \{f(\mathbf{x}) : \mathbf{x} \in X\} \subset F$. The *image* of f is $f(E)$, namely

$$\operatorname{im} f = \{f(\mathbf{x}) \in F : \mathbf{x} \in E\} \subset F.$$

The *inverse image* of a subset $Y \subset F$ is

$$f^{-1}(Y) = \{\mathbf{x} \in E : f(\mathbf{x}) \in Y\} \subset E.$$

Note that this definition does not require f to be invertible, but when it is, the inverse image of Y is also the image of Y by the inverse of f so that the notation is coherent.

Theorem. *Let $f : E \to F$ be a linear map. Then:*

(a) *The inverse image $f^{-1}(W)$ of any vector subspace $W \subset F$ is a vector subspace of E*

(b) *The image $f(V)$ of any vector subspace $V \subset E$ is a subspace of F.*

PROOF. (a) Let us consider a subspace W of F. If $\mathbf{x}, \mathbf{y} \in f^{-1}(W)$, namely $f(\mathbf{x}), f(\mathbf{y}) \in W$, then

$$f(a\mathbf{x} + \mathbf{y}) = af(\mathbf{x}) + f(\mathbf{y}) \in W.$$

This means that $a\mathbf{x} + \mathbf{y} \in f^{-1}(W)$, hence proves the first assertion.
(b) Let now V be any subspace of E. Take two elements of $f(V)$: They can be written in the form $f(\mathbf{x})$ and $f(\mathbf{y})$ for some $\mathbf{x}, \mathbf{y} \in V$. Then

$$af(\mathbf{x}) + f(\mathbf{y}) = f(a\mathbf{x} + \mathbf{y}) \in f(V)$$

is in the image of V. Hence $f(V)$ is a subspace of F. ■

Corollary 1. *For any linear map $f : E \to F$, $f^{-1}(0)$ is a subspace of E and $\operatorname{im} f = f(E)$ is a subspace of F.* ■

Corollary 2. *Consider two linear maps $f, g : E \to F$. Then the* coincidence subset

$$\{\mathbf{x} \in E : f(\mathbf{x}) = g(\mathbf{x})\}$$

is a vector subspace of E.

PROOF. The coincidence subset of f and g consists precisely of the $\mathbf{x} \in E$ such that

$$(f - g)(\mathbf{x}) = f(\mathbf{x}) - g(\mathbf{x}) = 0,$$

hence is the inverse image of the subspace $\{0\}$ by the linear map $f - g$. ■

Definitions. *Let $f : E \to F$ be a linear map. The* kernel *of f is the subspace $\ker f = f^{-1}(0)$ of E. The* image *of f is the subspace $f(E)$ of F. The* nullity *of f is the dimension (when finite) of its kernel. The* rank *of f is the dimension (when finite) of its image.*

If a linear map $f : E \to F$ is injective, then $0 \in E$ is the only element having image $0 \in F$, hence $\ker f = \{0\}$. There is a converse.

Proposition. *For a linear map f between vector spaces,*

$$\ker f = \{0\} \iff f \textit{ injective.}$$

PROOF. Quite generally, for any linear map f

$$\begin{aligned} f(\mathbf{x}) = f(\mathbf{y}) \quad &\iff \quad \underbrace{f(\mathbf{x}) - f(\mathbf{y})}_{f(\mathbf{x}-\mathbf{y})} = 0 \\ &\iff \quad \mathbf{x} - \mathbf{y} \in \ker f. \end{aligned}$$

If $\ker f = \{0\}$, we thus have

$$f(\mathbf{x}) = f(\mathbf{y}) \iff \mathbf{x} - \mathbf{y} = 0 \iff \mathbf{x} = \mathbf{y}.$$

This proves that f is injective. ■

The general form of the *basic principle* of linear algebra (Sec. 1.3.2) may now be formulated as follows. If $f : E \to F$ is linear, and $\mathbf{b} \in F$ is given, there is a solution $\mathbf{x} \in E$ of $f(\mathbf{x}) = \mathbf{b}$ precisely when $\mathbf{b} \in \operatorname{im} f$. When this is so, if $\mathbf{p}$ is a particular solution, namely $f(\mathbf{p}) = \mathbf{b}$, for any solution $\mathbf{x}$, we have

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{b} = f(\mathbf{p}), \\ f(\mathbf{x}-\mathbf{p}) &= f(\mathbf{x}) - f(\mathbf{p}) = 0, \\ \mathbf{x}-\mathbf{p} &\in \ker f. \end{aligned}$$

Hence any solution is the sum of the particular solution $\mathbf{p}$ and an element in the kernel of f:

$$\mathbf{x} = \mathbf{p} + \mathbf{x}_0 \quad (\mathbf{x}_0 \in \ker f).$$

4.2.2 How to Construct Linear Maps

Here is a general method for the construction of linear maps.

Proposition 1. *Let $(\mathbf{a}_i)_{i\in I}$ be a linearly independent subset of the vector space E. For any choice of a family $(\mathbf{b}_i)_{i\in I}$ in a vector space F, there is a linear map $f : E \to F$ such that $f(\mathbf{a}_i) = \mathbf{b}_i$ for all $i \in I$.*

PROOF. It is possible to enlarge the linearly independent family $(\mathbf{a}_i)_{i\in I}$ into a basis $(\mathbf{a}_i)_{i\in J}$ of E (if E is infinite dimensional, this is the mathematical credo). Now each element $\mathbf{x} \in E$ has a unique expression $\mathbf{x} = \sum_{i\in J} x_i\mathbf{a}_i$ as a linear combination of the elements of this basis. The components x_i are well-defined for all indices $i \in J$. We define $f(\mathbf{x}) = \sum_{i\in I} x_i\mathbf{b}_i$, forgetting the components x_i for $i \in J - I$. This map is linear. For example, one can check that it is homogeneous by simply observing that the components of $c\mathbf{x} = c\sum_{i\in J} x_i\mathbf{a}_i = \sum_{i\in J} cx_i\mathbf{a}_i$ are the scalars cx_i, whence

$$f(c\mathbf{x}) = \sum_{i\in I} cx_i\mathbf{a}_i = c\sum_{i\in I} x_i\mathbf{a}_i = cf(\mathbf{x}).$$

A similar verification shows the additivity of f. ■

Proposition 2. *Let $f, g : E \to F$ be two linear maps. If f and g agree on a subset $S \subset E$, then they agree on the linear span $\mathcal{L}(S)$. If f and g agree on a set of generators of E, then $f = g$.*

PROOF. The coincidence set of f and g is equal to the subspace $\ker(f - g)$ so that both statements follow. ■

The preceding argument may be presented in the following equivalent, but more computational way. For any $\mathbf{x} \in E$ in the linear span of a family $(\mathbf{a}_i)_{i\in I}$, there is (at least) one representation $\mathbf{x} = \sum_i x_i\mathbf{a}_i$. Hence we obtain

$$f(\mathbf{x}) = f(\Sigma x_i\mathbf{a}_i) = \Sigma_i x_i f(\mathbf{a}_i) = \Sigma_i x_i g(\mathbf{a}_i) = g(\Sigma x_i\mathbf{a}_i) = g(\mathbf{x}).$$

Due to its fundamental importance, we formulate explicitly a statement based on the preceding two propositions.

Proposition 3. Let $(\mathbf{e}_i)_{i\in I}$ *be a basis of the vector space* E. *For any choice of family* $(\mathbf{b}_i)_{i\in I}$ *in a vector space* F, *there is one and only one linear map* $f : E \to F$ *such that* $f(\mathbf{e}_i) = \mathbf{b}_i$ *for all* $i \in I$. ■

With the assumptions of Proposition 3, the linear map $f : \mathbf{e}_i \mapsto \mathbf{b}_i$ is

injective when the family $(\mathbf{b}_i)$ is *linearly independent* in F,

surjective when the family $(\mathbf{b}_i)$ *generates* F.

Comment. Let $\mathbf{e}_1, \ldots, \mathbf{e}_n$ denote the canonical basis of $\mathbf{R}^n$. Then any linear map $f : \mathbf{R}^n \to F$ is completely determined by the images $\mathbf{b}_i = f(\mathbf{e}_i)$ of the basis vectors, and these images may be prescribed arbitrarily. Here is the formula giving f explicitly

$$\mathbf{x} = \sum_{1\leqslant i\leqslant n} x_i\mathbf{e}_i \longmapsto f(\mathbf{x}) = \sum_{1\leqslant i\leqslant n} x_i\mathbf{b}_i.$$

We may write this in matrix form

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \longmapsto x_1\mathbf{b}_1 + \cdots + x_n\mathbf{b}_n = (\mathbf{b}_1 \ \ldots \ \mathbf{b}_n)\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Thus the row of vectors $(\mathbf{b}_1 \ \ldots \ \mathbf{b}_n)$ appears as the block matrix representing f.

4.2.3 Matrix Description of Linear Maps

The classical case of matrix description is the following.

Theorem. Let m *and* n *be two positive integers. Then any linear map* $f : \mathbf{R}^n \to \mathbf{R}^m$ *is given by a matrix multiplication: There is a unique* $m \times n$ *matrix* A *such that* $f(\mathbf{x}) = A\mathbf{x}$ *for all* $\mathbf{x} \in \mathbf{R}^n$.

PROOF. Let $(\mathbf{e}_j)_{1\leqslant j\leqslant n}$ denote the canonical basis of $\mathbf{R}^n$. Define the matrix A by

$$j\text{th column of } \mathbf{A} \ = f(\mathbf{e}_j).$$

Then consider the linear map g produced by matrix multiplication $g(\mathbf{x}) = A\mathbf{x}$ ($\mathbf{x} \in \mathbf{R}^n$). Then f and g agree on the generating set $(\mathbf{e}_j)$, hence $f = g$ by the Proposition 2 of the preceding subsection. ■

The matrix description of a linear map is a kind of photo of f depending on the viewpoint, here represented by a choice of bases. The problem of determining how the matrix description varies under a change of bases will be undertaken below (see Sec. 4.4.4).

Interpretation. Any linear plant (Sec. 4.1.1), or "black box" having a finite number of inputs and outputs, has an action given by a matrix product. Hence without any knowledge of the internal structure of the box, its function may be described by a matrix multiplication. The behavior of the box is fully determined when we only know the outputs corresponding to the basic inputs: All $s_i = 0$ except one $s_j = 1$ (for all possible values of j).

Remarks. (1) The statement of the preceding theorem is also valid when $m = 0$ and/or $n = 0$ if it is correctly interpreted, namely if the following convention is made. Each matrix consists of

$$\begin{cases} \text{its } \textit{size } m \times n \text{ (a pair of natural integers),} \\ \text{its } \textit{entries} \text{ (an array of numbers).} \end{cases}$$

Two matrices are *equal* only when they have the *same size and* the *same entries.* For example, a matrix of size $0 \times n$ has no entries, but the two matrices of sizes $0 \times n$ and $0 \times m$ (having no entries either) are equal only when $n = m$. The matrices $O_{m\times 0}$ and $O_{0\times n}$ can be multiplied, resulting in the zero matrix $O_{m\times n}$ of size $m \times n$: The empty "row by column" sums produce zeros at all relevant places. This product corresponds to the factorization $\mathbf{R}^n \to \{0\} \to \mathbf{R}^m$ of the trivial 0 map.

(2) Even without finiteness assumption, any linear map has a *generalized matrix* description. If $(\mathbf{e}_j)_{j\in J}$ is a basis of a vector space E, while $(\varepsilon_i)_{i\in I}$ is a basis of a vector space F, any linear map $f : E \to F$ is described by a generalized matrix A having entries a_{ij} such that

$$\begin{aligned} f(\mathbf{e}_j) &= \sum_{i\in I} a_{ij}\varepsilon_i, \\ f\Big(\sum_{j\in J} x_j\mathbf{e}_j\Big) &= \sum_{j\in J} x_j\Big(\sum_{i\in I} a_{ij}\varepsilon_i\Big). \end{aligned}$$

Since by definition there are only finitely many nonzero coefficients in these sums, they may be computed in any order, and we can write the result in the form $f(\mathbf{x}) = \mathbf{y} = \Sigma_i y_i\varepsilon_i$. The components

$$y_i = \sum_j a_{ij}x_j$$

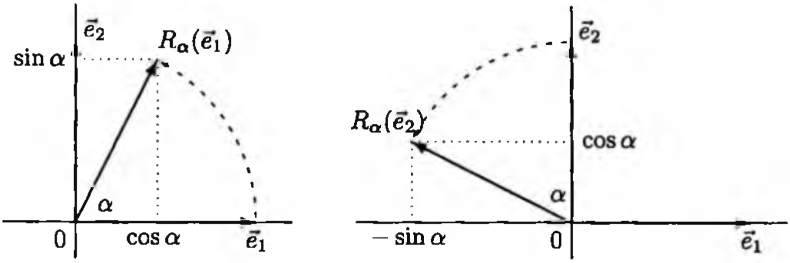
are given by a formal "row by column" matrix multiplication just as in the finite-dimensional case. The matrix of f appears as a generalized array of scalars

$$\mathrm{Mat}_{e,\varepsilon}(f) = (a_{ij})_{(i,j)\in I\times J}$$

depending on the choices of bases of E and F. In block form, the matrix of f is the row $(f(\mathbf{e}_j))_{j\in J}$ depending only on the choice of a basis in E. In the finite-dimensional case, we recover an array of size $m \times n$, where m is the cardinality of I and n the cardinality of J.

Application.

Let us consider a rotation of an angle α in the usual Cartesian plane $\mathbf{R}^2$. This is obviously a linear map $R_\alpha : \mathbf{R}^2 \to \mathbf{R}^2$. Hence it has a matrix description (in the canonical basis) with a 2×2 matrix having for columns the images of the basis vectors. Here are pictures



showing

$$R_\alpha(\vec{e}_1) = \begin{pmatrix} \cos\alpha \\ \sin\alpha \end{pmatrix}, \quad R_\alpha(\vec{e}_2) = \begin{pmatrix} -\sin\alpha \\ \cos\alpha \end{pmatrix}.$$

Hence the matrix M_α representing this rotation is

$$M_\alpha = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}.$$

Since the composition of two rotations is again a rotation, $R_{\alpha+\beta} = R_\alpha \circ R_\beta$, we see that

$$\begin{pmatrix} \cos(\alpha+\beta) & -\sin(\alpha+\beta) \\ \sin(\alpha+\beta) & \cos(\alpha+\beta) \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix} \begin{pmatrix} \cos\beta & -\sin\beta \\ \sin\beta & \cos\beta \end{pmatrix}$$

$$= \begin{pmatrix} \cos\alpha\cos\beta - \sin\alpha\sin\beta & -\sin\alpha\cos\beta - \cos\alpha\sin\beta \\ \sin\alpha\cos\beta + \cos\alpha\sin\beta & \cos\alpha\cos\beta - \sin\alpha\sin\beta \end{pmatrix}.$$

Comparing the first columns, we recover the addition formulas

$$\begin{cases} \cos(\alpha+\beta) &= \cos\alpha\cos\beta - \sin\alpha\sin\beta \\ \sin(\alpha+\beta) &= \sin\alpha\cos\beta + \cos\alpha\sin\beta. \end{cases}$$

For example, when $\beta = \alpha$

$$\begin{pmatrix} \cos 2\alpha & -\sin 2\alpha \\ \sin 2\alpha & \cos 2\alpha \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}^2$$

$$= \begin{pmatrix} \cos^2\alpha - \sin^2\alpha & -2\sin\alpha\cos\alpha \\ 2\sin\alpha\cos\alpha & \cos^2\alpha - \sin^2\alpha \end{pmatrix}.$$

We also recognize the duplication formulas

$$\begin{cases} \cos 2\alpha &= \cos^2\alpha - \sin^2\alpha \\ \sin 2\alpha &= 2\sin\alpha\cos\alpha. \end{cases}$$

More generally, from $R_{n\alpha} = R_\alpha^n$, we infer

$$\begin{pmatrix} \cos n\alpha & -\sin n\alpha \\ \sin n\alpha & \cos n\alpha \end{pmatrix} = \begin{pmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{pmatrix}^n.$$

4.3 The Dimension Theorem for Linear Maps

Let us examine more closely the situation for linear maps defined on a finite-dimensional vector space.

4.3.1 The Rank-Nullity Theorem

Here is a first formulation of one of the main results of linear algebra.

Theorem. Let E and F be two vector spaces. Assume that E is finite dimensional. Then for any linear map $f : E \to F$, the image of f is finite dimensional and

$$\dim\ker f + \dim\operatorname{im} f = \dim E.$$

PROOF. Put $k = \dim\ker f \leqslant n = \dim E$ and take a basis $(\mathbf{e}_1, \ldots, \mathbf{e}_k)$ of this kernel. It is possible to complete this linearly independent subset into a basis of E, say by adding elements $\mathbf{e}_{k+1}, \ldots, \mathbf{e}_n$. I claim that

$f(\mathbf{e}_{k+1}), \ldots, f(\mathbf{e}_n)$ *is a* basis *of the image* $f(E)$.

(*a*) If $\mathbf{y} \in f(E)$, we can find an element $\mathbf{x} \in E$ such that $\mathbf{y} = f(\mathbf{x})$. Since the $\mathbf{e}_i$ $(1 \leqslant i \leqslant n)$ form a basis of E, there is a (unique) representation $\mathbf{x} = \sum x_j \mathbf{e}_j$ so that $f(\mathbf{x}) = \sum x_j f(\mathbf{e}_j)$. But for $j \leqslant k$ the images $f(\mathbf{e}_j)$ are zero so that

$$\mathbf{y} = f(\mathbf{x}) = \textstyle\sum_{j>k} x_j f(\mathbf{e}_j).$$

This proves that the $f(\mathbf{e}_j)$ $(j > k)$ *generate* the image of f.
(*b*) Take any linear combination of the form

$$a_1 f(\mathbf{e}_{k+1}) + \cdots + a_{n-k} f(\mathbf{e}_n) = 0.$$

We see that

$$f(a_1\mathbf{e}_{k+1} + \cdots + a_{n-k}\mathbf{e}_n) = 0,$$

namely $a_1\mathbf{e}_{k+1} + \cdots + a_{n-k}\mathbf{e}_n \in \ker f$. By assumption, it is possible to write

$$a_1\mathbf{e}_{k+1} + \cdots + a_{n-k}\mathbf{e}_n = b_1\mathbf{e}_1 + \cdots + b_k\mathbf{e}_k \in \ker f,$$

whence a linear dependence relation

$$-b_1\mathbf{e}_1 - \cdots - b_k\mathbf{e}_k + a_1\mathbf{e}_{k+1} + \cdots + a_{n-k}\mathbf{e}_n = 0.$$

By the independence assumption, all coefficients in this relation are zero, and in particular

$$a_1 = 0, \ldots, a_{n-k} = 0.$$

Hence $f(\mathbf{e}_{k+1}), \ldots, f(\mathbf{e}_n)$ are *independent.*
Now (*a*) and (*b*) prove that the $f(\mathbf{e}_j)$ $(j > k)$ form a basis of the image of f: This space has dimension $n - k$, and

$$\dim\ker f + \dim\operatorname{im} f = k + (n - k) = n = \dim E. \qquad \blacksquare$$

Examples. (1) Consider the *vertical projection* onto the xy-plane in the usual 3-dimensional space

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \longmapsto \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} \quad : \mathbf{R}^3 \to \mathbf{R}^3.$$

The kernel of this linear map is the vertical axis, a 1-dimensional subspace, while its image is the 2-dimensional horizontal plane

$$1 + 2 = 3 = \dim \mathbf{R}^3.$$

(2) The derivation operator D, restricted to polynomials of degrees $\leqslant n$ has for kernel the constants (multiples of the constant 1). This kernel has dimension 1. The image of D consists of the polynomials of degrees $< n$, having basis $1 = x^0$, $x, \ldots, x^{n-1}$, hence of dimension n. The sum $1+n$ is indeed the dimension of the space of polynomials having degrees $\leqslant n$.

(3) Let $(a, b, c) \neq (0, 0, 0)$ and consider the linear form $\varphi : \mathbf{R}^3 \to \mathbf{R}$ defined by

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} \longmapsto ax + by + cz.$$

The kernel of φ is the subspace consisting of the solutions of the homogeneous equation $ax+by+cz = 0$. This homogeneous system consists of a single equation in three variables, hence has rank 1 since a, b, and c are not all 0. There is only one pivot variable, so that this space has dimension 2. It is a homogeneous plane (it contains the origin). More generally, when V is a finite-dimensional space and $\varphi : V \to \mathbf{R}$ is a nonzero linear form, then the image $\varphi(V) = \mathbf{R}$ has dimension 1 so that $\ker \varphi$ has dimension $\dim V - 1$. By analogy with the 3-dimensional case $V = \mathbf{R}^3$, we say that $\ker \varphi$ is a *hyperplane* in V.

4.3.2 Row-Rank versus Column-Rank

The rank-nullity theorem $\dim \ker f + \dim \operatorname{im} f = \dim E$ has important consequences. To formulate a particularly important one, we introduce a definition.

Definition. *The* transpose *of a matrix $A = (a_{ij})$ is the matrix $B = (b_{ij})$ having for rows the columns of A, namely $b_{ij} = a_{ji}$.*

If A has size $m \times n$, then B has size $n \times m$. Transposition is a *mirror symmetry* in the sense that if it is performed twice in succession, the original matrix is recovered. We denote by ${}^tA$ the transpose of A. Here is this symmetry

$$A = \begin{pmatrix} a & b & \\ a' & b' & \vdots \\ \vdots & \vdots & \end{pmatrix} \quad \longleftrightarrow \quad {}^tA = \begin{pmatrix} a & a' & \cdots \\ b & b' & \cdots \\ & \cdots & \end{pmatrix}.$$

For example, the transpose of a lower-triangular matrix is an upper-triangular matrix. In particular, with the notation introduced before (Sec. 3.2.1) ${}^tE_{ij} = E_{ji}$.

Theorem. The rank of a matrix A is equal to the rank of its transpose ${}^tA$.

PROOF. Consider the linear map $f : \mathbf{x} \mapsto A\mathbf{x}$ produced by left multiplication by A, so that $\ker f$ is the space of solutions of the homogeneous system $A\mathbf{x} = \mathbf{0}$. The dimension of this space is the number of free variables: $\dim\ker f = n - r$. Hence

$$\dim \operatorname{im} f = n - \dim\ker f = n - (n - r) = r.$$

Now the image of f consists precisely of the elements

$$f(\mathbf{x}) = f\left(\textstyle\sum_j x_j\mathbf{e}_j\right) = \textstyle\sum_j x_j f(\mathbf{e}_j),$$

namely

$$\operatorname{im} f = \mathcal{L}(f(\mathbf{e}_1), \ldots, f(\mathbf{e}_n)) = \mathcal{L}(\text{columns of } A).$$

If we define the column-rank of the matrix A by

$$s = \dim\mathcal{L}(\text{columns of } A) = \dim\mathcal{L}(\text{rows of } {}^tA),$$

we just proved $r = s$. ∎

Corollary. Let A be an $m \times n$ matrix. Then:

$$\begin{array}{rcl} A : \mathbf{R}^n \longrightarrow \mathbf{R}^m \textit{ surjective} & \Longleftrightarrow & {}^tA : \mathbf{R}^m \longrightarrow \mathbf{R}^n \textit{ injective} \\ A : \mathbf{R}^n \longrightarrow \mathbf{R}^m \textit{ injective} & \Longleftrightarrow & {}^tA : \mathbf{R}^m \longrightarrow \mathbf{R}^n \textit{ surjective.} \end{array}$$

PROOF. The rank of A is the dimension of its image, so that

$$\text{rank of } A \text{ is } m \quad \Longleftrightarrow \quad A \text{ surjective.}$$

By the rank-nullity theorem

$$\text{rank of } {}^tA \text{ is } m \quad \Longleftrightarrow \quad \underbrace{\dim\ker {}^tA = 0}_{\text{namely } \ker {}^tA = \{0\}} \quad \Longleftrightarrow \quad {}^tA \text{ injective.}$$

Hence the first statement follows from the equality of the ranks of A and of ${}^tA$. The second statement is obtained by replacing A by ${}^tA$. ∎

The equality row-rank(A) = column-rank(A) is not obvious, even for square matrices A. Since we consider it as a crucial result of linear algebra, let us rephrase its proof in a slightly different way. The row operations have been introduced in such a way as to preserve the set of solutions of the homogeneous system $Ax = 0$

row operations leave $\ker A$ *fixed.*

This basic property is also obvious in the interpretation of row operations as left multiplication by invertible matrices

$$\begin{aligned} A \sim B \quad &\Longleftrightarrow \quad B = EA \quad (E \text{ invertible}) \\ &\Longrightarrow \quad \ker B = \ker EA = \ker A. \end{aligned}$$

But *the image of A varies* when row operations are performed. However, the dimension relation: $\dim\ker f + \dim\operatorname{im} f = \dim E$, forces *the dimension of the image to remain constant* under row operations, since they leave $\dim\ker f$ fixed. If row operations carry A into an echelon form U, the images of A and of U have the same dimension. But a basis of the image of U consists of its pivot columns, so that $\dim\operatorname{im} U = r$.

4.3.3 Application: Invertible Matrices

Another important application of the rank-nullity theorem for a linear map $f : E \to E$ of a finite-dimensional space E *into itself* is

$$\underbrace{f \text{ injective}}_{\dim\ker f = 0} \quad \Longleftrightarrow \quad \underbrace{f \text{ surjective}}_{\dim\operatorname{im} f = \dim E} .$$

For the linear map $\mathbf{x} \mapsto A\mathbf{x} : \mathbf{R}^n \to \mathbf{R}^n$ produced by a square matrix A of size $n \times n$, it shows that A is injective when $n - r = \dim\ker A = 0$, hence when the rank of A is maximal: $r = n$.

Theorem. *For a square matrix A, the following conditions are equivalent:*

(*i*) *A is injective*
(*ii*) *A is surjective*
(*iii*) *A is bijective*
(*iv*) *A is invertible*
(*v*) *A has maximal rank $r = n$*
(*vi*) *A is row-equivalent to the identity.* ■

In particular, if A is left invertible, say $BA = I_n$, then A is injective, hence *invertible*. There is a square matrix C such that $AC = I_n$. By associativity of the matrix product

$$B = BI_n = B(AC) = (BA)C = I_nC = C,$$

so that left and right inverse coincide (another independent proof of this result may be found in the appendix to this chapter).

This situation should remind us of similar properties for maps of a *finite set* S into itself. With infinite sets, these equivalences fail (see Sec. 4.5.1). The above equivalences also fail for linear maps of an infinite-dimensional vector space into itself.

Let us now give a simple condition implying invertibility of a matrix. It is based on the existence of an absolute value on the scalars, namely a map $a \mapsto |a|$ having the properties

$$\begin{array}{l} |a| \geqslant 0 \textit{ for all scalars } a, \quad |a| = 0 \textit{ only for } a = 0, \\ |ab| = |a||b|, \quad |a+b| \leqslant |a| + |b| \quad (\textit{Triangle Inequality}). \end{array}$$

***Theorem** (Gershgorin). Let $A = (a_{ij})_{1 \leqslant i,j \leqslant n}$ be a square matrix satisfying*

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \quad (1 \leqslant i \leqslant n).$$

Then A is invertible.

PROOF. It is enough to prove that A is injective. Take any $\mathbf{x} = (x_j) \in \ker A$

$$\textstyle\sum_{1 \leqslant j \leqslant n} a_{ij} x_j = 0 \quad (1 \leqslant i \leqslant n).$$

Choose an index i such that $|x_i| = \max_j |x_j|$. We have

$$\begin{array}{rcl} -a_{ii} x_i & = & \sum_{j \neq i} a_{ij} x_j, \\ |a_{ii} x_i| & = & \left|\sum_{j \neq i} a_{ij} x_j\right|, \\ |a_{ii}|\,|x_i| & \leqslant & \sum_{j \neq i} |a_{ij}||x_j| \\ & \leqslant & |x_i| \sum_{j \neq i} |a_{ij}|. \end{array}$$

Hence we see that

$$\left(|a_{ii}| - \textstyle\sum_{j \neq i} |a_{ij}|\right) |x_i| \leqslant 0.$$

But by assumption we have $|a_{ii}| - \sum_{j \neq i} |a_{ij}| > 0$, so that $|x_i| = 0$. Since this component $|x_i|$ is maximal, all other ones are also zero, and $\mathbf{x} = 0$. ∎

Remark. The preceding statement uses the absolute value of the real field. As such, it cannot be formulated for *any* field of scalars. But it is also valid if scalars are taken in the complex field (see Sec. 11.2.2).

4.4 Isomorphisms

We have already studied invertible square matrices. Let us generalize the results obtained in the general context of vector spaces.

4.4.1 Generalities

***Definition.** An* isomorphism *is a linear bijective map between two vector spaces. If there is an isomorphism between two vector spaces, they are called* isomorphic.

Note that if $f : E \to F$ is an isomorphism, then the inverse $f^{-1} : F \to E$ is linear, hence also an isomorphism.

Proposition. *Two finite-dimensional vector spaces E and F are isomorphic precisely when they have the same dimension.*

PROOF. Take any isomorphism $f : E \to F$. Since f is injective, $\ker f = \{0\}$, $\dim \ker f = 0$; since f is surjective, $\operatorname{im} f = F$, $\dim \operatorname{im} f = \dim F$. Hence

$$\dim E = \dim \operatorname{im} f + \dim \ker f = \dim F.$$

Conversely, if E and F have the same dimension n, take bases $(\mathbf{e}_i)$ of E, $(\mathbf{f}_i)$ of F: Since they have the same number of elements by assumption, we may index them by the same index set $1 \leqslant i \leqslant n$. The mapping

$$\mathbf{x} = \textstyle\sum_i x_i \mathbf{e}_i \longmapsto \mathbf{y} = \sum_i x_i \mathbf{f}_i$$

defines an isomorphism between E and F. ∎

Examples. (1) We have identified linear homogeneous equations

$$a_1 x_1 + \cdots + a_n x_n = 0$$

with row vectors, namely matrices $(a_1, \ldots, a_n)$ of size $1 \times n$. Now we can say more precisely that linear homogeneous equations in n variables and row n-tuples constitute isomorphic vector spaces.

(2) The space $\mathbf{R}^4$ of 4-tuples and the space $\mathrm{Mat}_{2\times 2}(\mathbf{R})$ of 2×2 matrices have dimension 4. They are isomorphic vector spaces. Here is a natural isomorphism between them

$$\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} \longmapsto \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

(3) Row vectors and column vectors of the same length form isomorphic vector spaces. More generally, the transposition of matrices $A \mapsto {}^t\!A$ (Sec. 4.3.2) is an isomorphism $\mathrm{Mat}_{m\times n}(\mathbf{R}) \to \mathrm{Mat}_{n\times m}(\mathbf{R})$.

(4) A rotation in the usual space $\mathbf{R}^3$ is an isomorphism.

Comment. The elements $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ generate a vector space E precisely when the linear map

$$\begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} \longmapsto \sum_{1 \leqslant i \leqslant m} x_i \mathbf{a}_i$$

from $\mathbf{R}^m$ to E is *surjective* (hence $m \geqslant \dim E$). The same elements are linearly independent precisely when this linear map is *injective* (hence $m \leqslant \dim E$). The elements $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ form a basis of E precisely when the linear map

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \longmapsto \sum_{1 \leqslant i \leqslant n} x_i \mathbf{e}_i$$

from $\mathbf{R}^n$ to E is *bijective*—an *isomorphism*—hence when $n = \dim E$.

4.4.2 Models of Finite-Dimensional Vector Spaces

We have already proved that two finite-dimensional vector spaces of dimension n are isomorphic. In particular, the vector spaces $\mathbf{R}^n$ appear as paradigms of finite-dimensional spaces. Due to its importance, we formulate and prove again this result.

Theorem. *Every finite-dimensional vector space E of dimension $n \geqslant 1$ is isomorphic to $\mathbf{R}^n$.*

Proof. Let us choose a basis $(\mathbf{v}_i)_{1\leqslant i\leqslant n}$ of the vector space E. Hence each element $\mathbf{x} \in E$ has a unique representation as a linear combination of the $\mathbf{v}_i$'s, say $\mathbf{x} = \sum_i x_i\mathbf{v}_i$. Let $f(\mathbf{x})$ denote the n-tuple formed by the components of $\mathbf{x}$:

$$f(\mathbf{x}) = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbf{R}^n.$$

This map is bijective by definition. It is linear and its inverse $\mathbf{R}^n \to E$ is

$$(x_i)_{1\leqslant i\leqslant n} \longmapsto \textstyle\sum_{1\leqslant i\leqslant n} x_i\mathbf{v}_i. \qquad \blacksquare$$

The matrix description of a linear map $f : E \to F$ between finite-dimensional vector spaces, is a consequence of choices of bases in both domain and target spaces. These choices are reflected by vertical isomorphisms

$$\begin{array}{ccc} E & \xrightarrow{f} & F \\ (\mathbf{e})\updownarrow & & \updownarrow(\varepsilon) \\ \mathbf{R}^n & \xrightarrow{A} & \mathbf{R}^m. \end{array}$$

At the level of elements, recalling that the jth column of the matrix A of f is the image of the jth basis vector, or of its components in the chosen basis $(\mathbf{w}_i)$ of the target space

$$f(\mathbf{v}_j) = \sum_{1\leqslant i\leqslant m} a_{ij}\mathbf{w}_i \quad (1 \leqslant j \leqslant n),$$

we find

$$\mathbf{x} = \sum_{1\leqslant j\leqslant n} x_j\mathbf{v}_j \longmapsto f(\mathbf{x}) = \sum_{1\leqslant j\leqslant n} x_j f(\mathbf{v}_j) = \sum_{1\leqslant j\leqslant n} x_j \sum_{1\leqslant i\leqslant m} a_{ij}\mathbf{w}_i.$$

For the components, we see that the correspondence is given by the matrix multiplication

$$(x_i) \longmapsto A(x_i) = \left(\textstyle\sum_j a_{ij}x_j\right).$$

4.4.3 Change of Basis: Components of Vectors

A one-dimensional vector space E is simply a line with a marked 0 on it. Any nonzero element e constitutes a basis of this space. We may see it as a choice of unit. Any $x \in E$ has the form $x = a\,e$, where a is the component of x in this basis. If we take another choice of unit, say $\varepsilon = se$ with $s \neq 0$,

$$x = a\,e = (a/s)\,se = (a/s)\,\varepsilon$$

so that the component of the same element x is a/s. When we take a larger unit, the components are smaller: They vary in the opposite way, and for this reason, the term *contravariance* is also often used. This simple observation will now be generalized in an n-dimensional space.

Let $(\mathbf{e}_1, \dots, \mathbf{e}_n)$ denote the canonical basis of $\mathbf{R}^n$. We are going to consider simultaneously another basis $(\varepsilon_1, \dots, \varepsilon_n)$ of the same space, say

$$\varepsilon_j = \begin{pmatrix} s_{1j} \\ \vdots \\ s_{nj} \end{pmatrix} = \sum_{1 \leqslant i \leqslant n} s_{ij}\mathbf{e}_i.$$

Let $\mathbf{x} = (x_i) \in \mathbf{R}^n$ be any n-tuple. Its components in the canonical basis are the scalars x_i. How do we determine its components in the other basis? They are defined by

$$\mathbf{x} = \sum_{1 \leqslant i \leqslant n} x_i \mathbf{e}_i = \sum_{1 \leqslant i \leqslant n} y_i \varepsilon_i.$$

In other words, we are interested in the transformation $(x_i) \mapsto (y_i)$. This is certainly a linear operation, and in fact a bijective one, hence an isomorphism. A particularly simple case occurs for the n-tuples $\sum_{1 \leqslant i \leqslant n} s_{ij}\mathbf{e}_j = \varepsilon_j$ where the new components α_{ij} are all 0 except the jth equal to 1. In other words, the bijection we are looking for is determined by

$$\begin{pmatrix} s_{1j} \\ \vdots \\ s_{jj} \\ \vdots \\ s_{nj} \end{pmatrix} \longmapsto \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \leftarrow j\text{ th place.}$$

Thus the inverse is determined by

$$\mathbf{e}_j \longmapsto j\text{th column of } S,$$

hence is S itself. We infer that the required isomorphism is S^{-1} and we have solved our problem:

$$(y_i) = S^{-1}(x_i).$$

Here is a diagram summing up the situation:

$$\begin{aligned}
[\mathbf{x}]_e &= (x_i)_{1\leqslant i\leqslant n}: \quad n\text{-tuple of components of } \mathbf{x} \text{ in the basis } (\mathbf{e}_i)\\
[\mathbf{x}]_\varepsilon &= (y_i)_{1\leqslant i\leqslant n}: \quad n\text{-tuple of components of } \mathbf{x} \text{ in the basis } (\varepsilon_i).
\end{aligned}$$

We have established

$$[\mathbf{x}]_\varepsilon = S^{-1}[\mathbf{x}]_e, \quad [\mathbf{x}]_e = S[\mathbf{x}]_\varepsilon,$$

while by definition $S(\mathbf{e}_j) = (j\text{ th column of } S) = \varepsilon_j$. Note that the inverse of S—and not S itself—occurs in this transformation formula, just as in the one-dimensional case. The components of a vector are therefore also called *contravariant components* to recall this transformation property (and in particular to distinguish them from other—false components—to be defined later: see Sec. 11.6.2).

Let us show how the preceding formulas are written with block multiplication. First, the formulas

$$\varepsilon_j = \begin{pmatrix} s_{1j} \\ \vdots \\ s_{nj} \end{pmatrix} = \sum_{1\leqslant i\leqslant n} s_{ij}\mathbf{e}_j \quad (1 \leqslant j \leqslant n)$$

may simply be rewritten

$$(\varepsilon_1, \dots, \varepsilon_n) = (\mathbf{e}_1, \dots, \mathbf{e}_n)S.$$

Hence

$$\mathbf{x} = \sum x_i\mathbf{e}_i = (\mathbf{e}_1, \dots, \mathbf{e}_n)\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \underbrace{(\mathbf{e}_1, \dots, \mathbf{e}_n)S}_{(\varepsilon_1, \dots, \varepsilon_n)} \cdot S^{-1}\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Here we read the components in the ε-basis

$$\mathbf{x} = \sum y_i\varepsilon_i = (\varepsilon_1, \dots, \varepsilon_n)\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad \text{where} \quad \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = S^{-1}\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}.$$

Formally, the situation is similar to the one-dimensional case.

Finally, observe that since $(\mathbf{e}_i)$ is the canonical basis of $\mathbf{R}^n$, $(\mathbf{e}_1, \dots, \mathbf{e}_n)$ is the unit matrix so that we recover

$$S = (\varepsilon_1, \dots, \varepsilon_n) = \begin{pmatrix} s_{11} & \dots & s_{1n} \\ \vdots & & \vdots \\ s_{n1} & \dots & s_{nn} \end{pmatrix}.$$

4.4.4 Change of Basis: Matrices of Linear Maps

Let us consider a linear map $f : \mathbf{R}^n \to \mathbf{R}^n$. As we have seen, it is represented by a matrix, say $f(\mathbf{x}) = A\mathbf{x}$. Consider a new basis $(\varepsilon_i)_{1\leqslant i\leqslant n}$ of $\mathbf{R}^n$. What is the matrix of f in this new basis? Call

$$A = \mathrm{Mat}_e(f) = \mathrm{Mat}_{e,e}(f), \quad B = \mathrm{Mat}_\varepsilon(f) = \mathrm{Mat}_{\varepsilon,\varepsilon}(f).$$

Consider the following diagram:

$$\begin{tikzcd}
{[\mathbf{x}]_e} \arrow[r, mapsto, "A"] & {[f(\mathbf{x})]_e} \arrow[d, "S^{-1}"] \\
{[\mathbf{x}]_\varepsilon} \arrow[u, "S"] \arrow[r, mapsto, "B"] & {[f(\mathbf{x})]_\varepsilon}\,.
\end{tikzcd}$$

We infer that

$$B = S^{-1}AS.$$

Two matrices A and B related by such an equality are called *similar matrices.*

More generally, if $f : E \to F$ is a linear map, selecting an *input basis* $(\mathbf{e}_i)$ of E, an *output basis* (ε_j) of F, the matrix representation $A = \mathrm{Mat}_{e,\varepsilon}(f)$ is given by the following diagram:

$$\begin{tikzcd}
\mathbf{x} \arrow[r, mapsto, "f"] \arrow[d, leftrightarrow] & f(\mathbf{x}) \arrow[d, leftrightarrow] \\
{[\mathbf{x}]_e} \arrow[r, mapsto, "A"] & {[f(\mathbf{x})]_\varepsilon}\,.
\end{tikzcd}$$

4.4.5 The Trace of Square Matrices

By a change of basis, the appearance of a matrix changes a lot. However, the sum of the diagonal elements of a square matrix is *invariant.*

Definition. *Let $A = (a_{ij})$ be a square matrix of size $n \times n$. Then the sum $\operatorname{tr} A = \sum_{1\leqslant i\leqslant n} a_{ii}$ of the diagonal elements is called the* trace *of the matrix A.*

The trace is a linear form on the vector space of square matrices of given size $n \times n$: It takes scalar values.

Proposition. *Let A be a matrix of size $n \times m$, B a matrix of size $m \times n$, so that AB and BA are well-defined, square matrices of sizes $n \times n$ and $m \times m$ respectively. Then*

$$\operatorname{tr}(AB) = \operatorname{tr}(BA).$$

PROOF. The square matrix AB has size $n \times n$ and its diagonal entries are

$$(AB)_{ii} = a_{i1}b_{1i} + \cdots + a_{im}b_{mi} \quad (1 \leqslant i \leqslant n).$$

Hence

$$\begin{aligned}\operatorname{tr}(AB) &= \sum_{1\leqslant i\leqslant n} a_{i1}b_{1i} + \cdots + \sum_{1\leqslant i\leqslant n} a_{im}b_{mi}\\ &= \sum_{1\leqslant i\leqslant n} b_{1i}a_{i1} + \cdots + \sum_{1\leqslant i\leqslant n} b_{mi}a_{im}\\ &= (BA)_{11} + \cdots + (BA)_{mm} \;=\; \operatorname{tr}(BA),\end{aligned}$$

whence the result. ■

In particular, if $A = (a_1, \ldots, a_m)$ is a row and $B = {}^t(b_1, \ldots, b_m)$ is a column, the trace of the square matrix BA (of rank less than or equal to one) is equal to the scalar AB.

Corollary 1. *For square matrices A and B of the same size, we have*

$$\operatorname{tr}(AB) = \operatorname{tr}(BA).$$

■

We can now show that the trace of a matrix is *invariant* with respect to a coordinate change.

Corollary 2. *Let A be a square matrix, S an invertible matrix of the same size. Then* $\operatorname{tr}(S^{-1}AS) = \operatorname{tr} A$.

PROOF. Use the first corollary for the matrices $A_1 = S^{-1}$ and $A_2 = AS$. We obtain

$$\operatorname{tr}(S^{-1}AS) = \operatorname{tr}(A_1A_2) = \operatorname{tr}(A_2A_1) = \operatorname{tr}(ASS^{-1}) = \operatorname{tr} A.$$

■

Caution. The trace of a product *does* in general *depend* on the order of its factors: If we exchange the first two factors in a product of $n \geqslant 3$ factors

$$\operatorname{tr}(A_1A_2\cdots A_n) \neq \operatorname{tr}(A_2A_1\cdots A_n) \quad \text{(in general).}$$

The proposition only shows that a *circular permutation* does not change its value. For example, with three matrices,

$$\operatorname{tr}(ABC) = \operatorname{tr}(BCA) = \operatorname{tr}(CAB)$$

may differ from

$$\operatorname{tr}(BAC) = \operatorname{tr}(ACB) = \operatorname{tr}(CBA).$$

4.5 Appendix

4.5.1 Inverting Maps Between Sets

If $f : E \to F$ and $g : F \to G$ are two maps, then:

$$\begin{aligned} g \circ f : E \to F \to G \text{ injective} &\quad\Longrightarrow\quad f \text{ injective}\\ g \circ f : E \to F \to G \text{ surjective} &\quad\Longrightarrow\quad g \text{ surjective.}\end{aligned}$$

For example if $G = E$

$$g \circ f = id_E \quad \Longrightarrow \quad f \text{ injective and } g \text{ surjective.}$$

But in general, nothing more can be said. Here is an example. Consider the maps $f, g : \mathbf{N} \to \mathbf{N}$ defined by

$$f(n) = 2n, \quad g(2n) = g(2n+1) = n$$

(so that $g(m)$ is the integral part of $m/2$). Obviously $g(f(n)) = g(2n) = n$ hence

$$g \circ f = id.$$

However, neither f nor g is bijective: Neither is invertible. The left inverse g of f is not a right inverse of f. Here is a representation of this situation.

$$\mathbf{N} \xrightarrow{f} \boxed{\begin{array}{c} \nearrow \boxed{\text{even integers}} \searrow \\ \boxed{\text{odd integers}} \nearrow \end{array}} \xrightarrow{g} \mathbf{N}$$

But if a map $f : E \to F$ is both left and right invertible, then these inverses coincide by associativity of the composition of maps: If $g, h : F \to E$ are two maps such that $g \circ f = id_E$, $f \circ h = id_F$ then

$$g \circ f \circ h = \begin{cases} g \circ id_F &= g \\ id_E \circ h &= h. \end{cases}$$

Finally, if f and g are invertible, the composite $g \circ f$ is also invertible

$$\begin{aligned} g \circ f \quad &: \quad E \xrightarrow{f} F \xrightarrow{g} G, \\ (g \circ f)^{-1} \quad &: \quad E \xleftarrow{f^{-1}} F \xleftarrow{g^{-1}} G, \\ (g \circ f)^{-1} \quad &= \quad f^{-1} \circ g^{-1}. \end{aligned}$$

More generally, a composition of invertible maps is invertible and

$$(f_n \circ f_{n-1} \circ \cdots \circ f_1)^{-1} = f_1^{-1} \circ \cdots \circ f_{n-1}^{-1} \circ f_n^{-1}.$$

To undo a composition, start by undoing what has been done last: Just as you would with knitting!

4.5.2 Another Proof of Invertibility

Let us show how one may establish the main invertibility result without the rank theory.

Proposition. *Let E be a finite-dimensional space and let $f : E \to E$ be a linear map. If f is* left invertible, *then f is invertible.*

PROOF. Let $n = \dim E$, so that the dimension of the space of linear maps $E \to E$ is $N = n^2$ (This is obvious with the matrix representation of linear maps.) The $N+1$ linear maps

$$I_E,\ f,\ f^2, \ldots, f^N$$

cannot be linearly independent: Let

$$\sum_{i \leqslant N} a_i f^i = 0$$

be a nontrivial linear dependence relation. There is a smallest index j with nonzero coefficient

$$a_j f^j + a_{j+1} f^{j+1} + \cdots = 0 \quad (a_j \neq 0),$$

and multiplying by the inverse of a_j, we get a simpler relation

$$f^j (I_E + b_{j+1} f + \cdots) = 0.$$

Using the left inverse of f (j times), we deduce

$$I_E + b_{j+1} f + b_{j+2} f^2 + \cdots = 0.$$

Hence

$$I_E = f(-b_{j+1} - b_{j+2} f - \cdots),$$

showing that f has a right inverse and is thus invertible.

Example. Let N be a square matrix of size $n \times n$, satisfying $N^k = O$ for some integer $k \geqslant 1$. Then $I_n \pm N$ are invertible and for example

$$(I_n - N)^{-1} = I_n + N + N^2 + \cdots + N^{k-1}.$$

Indeed, the product of this finite sum with $I_n - N$ is

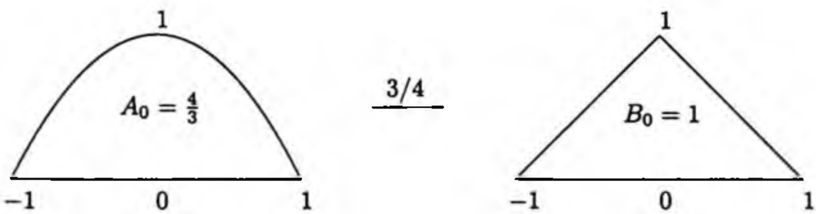
$$\begin{array}{rcl} I_n & + & N + N^2 + \cdots + N^{k-1} \\ & - & N - N^2 - \cdots - N^{k-1} - N^k \\ & = & I_n - N^k \\ & = & I_n. \end{array}$$

An illustration of the use of linearity

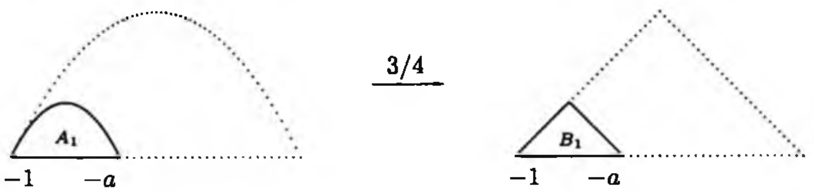
The area under the parabola $y = 1 - x^2$ between $x = -1$ and $x = 1$ is

$$\int_{-1}^{1} (1 - x^2)\,dx = 2\int_{0}^{1} (1 - x^2)\,dx = 2\Big[x - \frac{x^3}{3}\Big]_0^1 = 2(1 - 1/3) = 4/3.$$

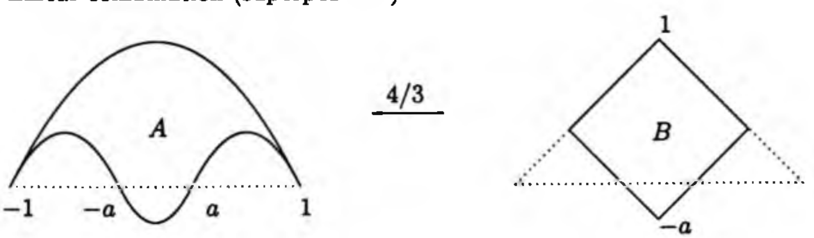
We can use this to compute the area of more sophisticated figures as follows.



Proportionality



Linear combination (superposition)



$A = A_0 - 2A_1 + A_2$ $\qquad$ $B = (1 + a)\frac{1+a}{2}$

Conclusion

$$A = \tfrac{2}{3}(1 + a)^2.$$

This sequence of figures is adapted from an article by R. Nelsen:

The Area of a Salinon: The Mathematical Magazine **75**, nb.2 (2002) p.130.

4.6 Exercises

1. Is the transformation between \$ and £ linear? Same question for Fahrenheit and Celsius degrees, Greenwich Mean Time and local time in New York, radians and degrees.

2. (*a*) Let $f : \mathbf{R}^3 \to \mathbf{R}^3$ be given by left multiplication by the matrix

$$A = \begin{pmatrix} a & a' & a'' \\ b & b' & b'' \\ c & c' & c'' \end{pmatrix}$$

in the canonical basis $\vec{e}_1, \vec{e}_2, \vec{e}_3$ of $\mathbf{R}^3$. What is the matrix of f in the basis $\vec{e}_3, \vec{e}_2, \vec{e}_1$?
(*b*) In the canonical basis of $\mathbf{R}^4$, a linear map f is described by the matrix

$$A = \begin{pmatrix} a & 0 & 0 & b \\ 0 & a & b & 0 \\ 0 & b & a & 0 \\ b & 0 & 0 & a \end{pmatrix}.$$

What is the matrix of f in the basis e_1, e_4, e_2, e_3? Is A similar to

$$B = \begin{pmatrix} b & 0 & 0 & a \\ 0 & b & a & 0 \\ 0 & a & b & 0 \\ a & 0 & 0 & b \end{pmatrix}?$$

(*c*) Let $f : \mathbf{R}^n \to \mathbf{R}^n$ be an invertible linear map. Compare the matrix A of f in a basis $(\mathbf{e}_i)_{1 \leqslant i \leqslant n}$ and the matrix B of f in the basis $(f(\mathbf{e}_i))_{1 \leqslant i \leqslant n}$. More generally, let $f : \mathbf{R}^n \to \mathbf{R}^n$ be a linear map, $g : \mathbf{R}^n \to \mathbf{R}^n$ be an invertible linear map such that $fg = gf$. Compare the matrix A of f in a basis $(\mathbf{e}_i)_{1 \leqslant i \leqslant n}$ and the matrix B of f in the basis $(g(\mathbf{e}_i))_{1 \leqslant i \leqslant n}$.

3. We consider the following maps of the space of polynomials Π into itself

$$\begin{aligned} p(x) &\longmapsto p(x-1), \\ p(x) &\longmapsto 1 + p(x-1), \\ p(x) &\longmapsto -p(x+2), \\ p(x) &\longmapsto 2p(x^3), \\ p(x) &\longmapsto p(x)^2. \end{aligned}$$

Which ones are linear?

4. Using the matrix description of rotations given in 4.2.3, show that $\cos n\alpha = T_n(\cos\alpha)$ $(n \geqslant 0)$ is a polynomial of degree n in $\cos\alpha$ (Chebyshev polynomials of the first kind). Show that $\sin(n+1)\alpha = \sin\alpha\, U_n(\cos\alpha)$ $(n \geqslant 0)$ with a polynomial U_n of degree n (Chebyshev polynomials of the second kind).

5. Consider the two matrices

$$A = \begin{pmatrix} 1 & 1/2 \\ 0 & 1/2 \end{pmatrix}, \quad B = \begin{pmatrix} 1/2 & 0 \\ 1/2 & 1 \end{pmatrix}.$$

Determine the image of the region $x \geqslant 0$, $y \geqslant 0$ by

$$A,\ B,\ AB,\ BA,\ ABA,\ A^2B,\ BAB.$$

Conclude (without computation) that all finite products

$$X_1 X_2 \cdots X_n,$$

where each factor X_i is equal to A or B, are different.

6. Consider the following three matrices

$$A = \begin{pmatrix} 1 & 1/2 & 1/2 \\ 0 & 1/2 & 0 \\ 0 & 0 & 1/2 \end{pmatrix}, \quad B = \begin{pmatrix} 1/2 & 0 & 0 \\ 1/2 & 1 & 1/2 \\ 0 & 0 & 1/2 \end{pmatrix}, \quad C = \begin{pmatrix} 1/2 & 0 & 0 \\ 0 & 1/2 & 0 \\ 1/2 & 1/2 & 1 \end{pmatrix}.$$

Let Δ be the basic triangle having vertices at the extremities of the canonical basis. What is the image of this triangle Δ by the linear maps (make pictures)

$$A,\quad B,\quad C,\quad AB,\quad BA,\quad A^2B,\quad ABC, \ldots ?$$

Conclude (without computation) that all finite products in A, B, and C are different.

7. Let $\Pi_{x,y}$ denote the space of polynomials in two variables x and y, and V the subspace of polynomials of total degree less than or equal to 3: V is spanned by the monomials $x^i y^j$ for integers i and j such that $0 \leqslant \deg(x^i y^j) = i + j \leqslant 3$.
(a) Give a basis of V: What is the dimension of this space? The linear maps $\partial/\partial x$, $\partial/\partial y$: $V \to V$ are defined by

$$\partial/\partial x(x^i y^j) = i x^{i-1} y^j, \quad \text{resp.} \quad \partial/\partial y(x^i y^j) = j x^i y^{j-1}$$

(by convention $ix^{i-1} = 0$ for $i = 0$, and $jy^{j-1} = 0$ for $j = 0$). Give the matrix of the linear map $F = \partial/\partial x - \partial/\partial y : V \to V$ in the chosen basis. What is the dimension of $\ker f$ and $\operatorname{im} f$?

Give the matrix of the linear map $D = x(\partial/\partial x) - y(\partial/\partial y) : V \to V$ (in the same basis).

(b) Let $\Delta = \partial^2/\partial^2 x + \partial^2/\partial^2 y$ be the *Laplace operator* in $\Pi_{x,y}$. Prove that the kernel of Δ is infinite dimensional: Its elements are the *harmonic polynomials*. Find a basis of $V \cap \ker \Delta$.

8. Let $\mathbf{M}_3$ denote the vector space of 3×3 matrices and let

$$M = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

We consider the following linear maps $\mathbf{M}_3 \to \mathbf{M}_3$

$$\begin{aligned} f &: A \longmapsto AM, \\ g &: A \longmapsto MAM. \end{aligned}$$

Give bases for their kernels. What are the dimensions of their images: Check the rank-nullity theorem for f and g.

9. Prove that the following matrices are invertible

$$\begin{pmatrix} 4 & 1 & 1 & 1 \\ 1 & 5 & 1 & -1 \\ 1 & 1 & -3 & 0 \\ 2 & -2 & 1 & 6 \end{pmatrix}, \quad \begin{pmatrix} 4 & 5 & 0 & 1 \\ 1 & 10 & -1 & 1 \\ 1 & 1 & 3 & 1 \\ 1 & -3 & 1 & 4 \end{pmatrix}.$$

10. Let E and F be two vector spaces. Prove that f is linear whenever its graph $\left\{\binom{x}{y} : y = f(x)\right\} \subset E \times F$ is a subspace of $E \times F$ (with its natural structure of vector space).

11. Let

$$0 \xrightarrow{f_0} V_1 \xrightarrow{f_1} V_2 \xrightarrow{f_2} \cdots \xrightarrow{f_{n-1}} V_n \xrightarrow{f_n} 0$$

be a sequence of linear maps with $\operatorname{im} f_i = \ker f_{i+1}$ $(0 \leqslant i < n)$. Prove

$$\sum_{1 \leqslant i \leqslant n} (-1)^i \dim V_i = 0.$$

12. Let

$$\begin{array}{ccc} V_1 & \xrightarrow{f_1} & V_2 \\ f_4 \uparrow & & \downarrow f_2 \\ V_4 & \xleftarrow{f_3} & V_3 \end{array}$$

be a diagram consisting of linear maps with $\operatorname{im} f_i = \ker f_{i+1}$ $(1 \leqslant i \leqslant 4$, where $f_5 = f_1))$. Prove $\dim V_1 + \dim V_3 = \dim V_2 + \dim V_4$. What happens in the particular case where $f_2 = 0$ and $f_4 = 0$?

13. Let E, F be two vector spaces, $\operatorname{Hom}(E, F)$ denote the set of linear maps $f : E \to F$, and $\operatorname{Sub}(E)$ the set of vector subspaces of E. Is the map

$$\begin{array}{ccc} (E, F) & \longrightarrow & \operatorname{Sub}(E) \\ f & \longmapsto & \ker f \end{array}$$

injective, surjective?

I may choose freely the images of the basis vectors!

Chapter 5

The Rank Theorem

The abstract language of linear algebra has allowed us to prove the invariance of the rank, as well as the equalities between row- and column-rank. This is the rank theory, which has far-reaching consequences: Some of them are presented in this chapter. On the other hand, we give further examples showing how the general concept of vector does enrich the mathematical description of natural phenomena:

- A *velocity* (vector) is more precise than a *speed* (scalar)
- A *colored* pixel is richer than a *grey* one
- A *pyramid of ages* contains more information than a bare *total population* figure.

5.1 More on Row- versus Column-Rank

We have already seen that the row-rank $r = \dim \mathcal{L}(\text{rows of } A)$ is the same as the column-rank $\dim \mathcal{L}(\text{columns of } A) = \dim \operatorname{im}(A)$. In other words, the ranks of a matrix A and its transpose ${}^tA$ are the same. More can be said.

5.1.1 Factorizations of a Matrix

Proposition. *Any matrix A of size $m \times n$ and rank r admits a factorization $A = ST$ where the size of S is $m \times r$ and the size of T is $r \times n$.*

PROOF. Proceeding with row operations, we can find an echelon form of the matrix A, say $A \sim U = EA$, where E is invertible of size $m \times m$

$$A = E^{-1}U = \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{pmatrix} \begin{pmatrix} * & \cdots & \cdots & * \\ 0 & & & * \\ 0 & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & 0 \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} * \\ 0 \end{matrix}}\right\} r \\ \left.\vphantom{\begin{matrix} 0 \\ 0 \end{matrix}}\right\} m-r \end{matrix}$$

In this matrix product, the last $m-r$ columns of E^{-1} may be ignored together with the last $m-r$ rows of U since these are identically zero. This produces a factorization $A = E_r^{-1}U_r$ of the desired form. ∎

Comment. A factorization of a matrix A of rank r and size $m \times n$ corresponds to a factorization

$$A : \mathbf{R}^n \xrightarrow{T} \mathbf{R}^r \xrightarrow{S} \mathbf{R}^m.$$

The preceding factorization may also be deduced as follows. Let us take any basis of the row space of A, say $\mathbf{t}_1, \dots, \mathbf{t}_r$. All rows may be expressed as linear combinations of these

$$\rho_i = s_{i1}\mathbf{t}_1 + \cdots + s_{ir}\mathbf{t}_r = \textstyle\sum_{1\leqslant k\leqslant r} s_{ik}\mathbf{t}_k.$$

Take the jth component of these row identities

$$a_{ij} = \rho_{ij} = s_{i1}t_{1j} + \cdots + s_{ir}t_{rj} = \textstyle\sum_{1\leqslant k\leqslant r} s_{ik}t_{kj}.$$

Here is a matrix factorization $A = ST$ with $S = (s_{ik})$ of size $m \times r$ and $T = (t_{kj})$ of size $r \times n$. Grouping the preceding equations in a column, we find

$$\mathbf{a}_j = t_{1j}\mathbf{s}_1 + \cdots + t_{rj}\mathbf{s}_r = \textstyle\sum_{1\leqslant k\leqslant r} \mathbf{s}_k t_{kj},$$

for the corresponding $m$tuples, and this shows that

$$\text{column-rank}\,(A) \leqslant r = \text{row-rank}\,(A).$$

Of course, the same result holds for the transpose of A

$$\underbrace{\text{column-rank}\,({}^tA)}_{\text{row-rank}\,(A)} \leqslant \underbrace{\text{row-rank}\,({}^tA)}_{\text{column-rank}\,(A)}.$$

This proves once more the equality of the ranks in question, and therefore we may simply speak of *rank* of a matrix without specifying which one is considered.

5.1.2 Low Rank Examples

Let us first consider the rank 1 case: All rows of A are proportional (all columns are also proportional). Such a matrix can be factorized into a product of a column matrix by a row matrix

$$A = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} (b_1 \dots b_n) = \begin{pmatrix} a_1b_1 & \dots & a_1b_n \\ \vdots & & \vdots \\ a_mb_1 & \dots & a_mb_n \end{pmatrix}.$$

Any matrix of this type has rank less than or equal to 1: The rank is 0 when all entries are 0. All columns of A are proportional to $\mathbf{a} = {}^t(a_1, \dots, a_m)$, hence the image of A consists of multiples of this vector. More precisely

$$A\mathbf{x} = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} (b_1 \dots b_n) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_1 \\ \vdots \\ a_m \end{pmatrix} \varphi(\mathbf{x}),$$

where $\varphi(\mathbf{x}) = b_1x_1 + \cdots + b_nx_n$. This shows that for any linear $A : \mathbf{R}^n \longrightarrow \mathbf{R}^m$ of rank less than or equal to 1, there is a linear form $\varphi : \mathbf{R}^n \to \mathbf{R}$ and a vector $\mathbf{a} \in \mathbf{R}^m$ such that $A\mathbf{x} = \mathbf{a}\,\varphi(\mathbf{x})$, namely $A = \mathbf{a}\,\varphi$.

Here are some examples of rank 2. Let x_i, y_j $(1 \leqslant i, j \leqslant n)$ be arbitrary real numbers and consider

$$A = (a_{ij}) \text{ with } a_{ij} = \cos(x_i - y_j).$$

I claim that the rank of A is less than or equal to 2. Indeed

$$\cos(x_i - y_j) = \cos x_i \cos y_j + \sin x_i \sin y_j,$$

whence $A = C + S$ with

$$C = (\cos x_i \cos y_j) = (a_i b_j), \quad S = (\sin x_i \sin y_j) = (c_i d_j).$$

By the preceding considerations, we see that $\operatorname{rank}(C) \leqslant 1$, $\operatorname{rank}(S) \leqslant 1$. Since the rank of a matrix is the dimension of its image, and each element of $\operatorname{im}(C+S)$ is the sum of an element in $\operatorname{im} C$ and an element in $\operatorname{im} S$, we have

$$\dim \operatorname{im}(C + S) \leqslant \dim \operatorname{im} C + \dim \operatorname{im} S \leqslant 2.$$

Another way of reaching the same conclusion consists in writing the matrix product

$$A = \begin{pmatrix} \cos x_1 & \sin x_1 \\ \cos x_2 & \sin x_2 \\ \vdots & \vdots \\ \cos x_n & \sin x_n \end{pmatrix} \begin{pmatrix} \cos y_1 & \cdots & \cos y_n \\ \sin y_1 & \cdots & \sin y_n \end{pmatrix},$$

corresponding to a factorization $A : \mathbf{R}^n \to \mathbf{R}^2 \to \mathbf{R}^n$. Hence the dimension of the image of the composite cannot be greater than the dimension of the image of the second map, having rank less than or equal to 2.

5.1.3 A Basis for the Column Space

Let A be a matrix of size $m \times n$ having columns $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbf{R}^m$, so that

$$\mathcal{L}(\text{columns of } A) = \mathcal{L}(\mathbf{a}_1, \ldots, \mathbf{a}_n) \subset \mathbf{R}^m.$$

Any linear dependence relation between columns

$$x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n = 0,$$

can be written

$$A\mathbf{x} = (\mathbf{a}_1 \ \ldots \ \mathbf{a}_n) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = 0,$$

namely corresponds to an element $\mathbf{x}$ in $\ker A$. Since *row operations preserve the kernel, they also preserve the linear relations between columns.* In particular, if $A \sim B$, any subset of independent columns of A corresponds to a subset of independent columns (same indices) of B. If the rank of A is r, we can take for B a row-reduced form U, with pivot columns having indices $j_1 < \cdots < j_r$. The columns of A of the same indices $j_1 < \cdots < j_r$ are independent, hence

> *The columns of A of indices $j_1 < \cdots < j_r$ constitute a* basis *of the r-dimensional space $\mathcal{L}$(columns of A).*

Let us say it slightly differently: Any row operation is given by a left multiplication $A \sim B = EA$, where E denotes an invertible matrix of size $m \times m$, hence an isomorphism $\mathbf{R}^m \to \mathbf{R}^m$. Since the columns of $B = EA$ are $E\mathbf{a}_j$ $(1 \leqslant j \leqslant n)$, we see that the left multiplication by E defines an isomorphism

$$\mathcal{L}(\text{columns of } A) = \mathcal{L}(\mathbf{a}_1, \ldots, \mathbf{a}_n) \xrightarrow{\sim} \mathcal{L}(E\mathbf{a}_1, \ldots, E\mathbf{a}_n) = \mathcal{L}(\text{columns of } B).$$

The row operations carry a basis of the column space of A onto a basis of the column space of U. Reversing the operations, any basis of the column space of U is mapped 1-1 onto a basis of the column space of A.

5.2 Direct Sum of Vector Spaces

5.2.1 Sum of Two Subspaces

Let V and W be two subspaces of a vector space E. We define their *sum* by

$$V + W = \{\mathbf{v} + \mathbf{w} : \ \mathbf{v} \in V, \ \mathbf{w} \in W\}.$$

This is the subspace of E generated by V and W. If A is a set of generators of V, B a set of generators of W, then $A \cup B$ generates $V + W$. If $\mathbf{v} \in V$, $\mathbf{w} \in W$, and $\mathbf{x} \in V \cap W$, then

$$\mathbf{v} + \mathbf{w} = \underbrace{(\mathbf{v} + \mathbf{x})}_{\mathbf{v}' \in V} + \underbrace{(\mathbf{w} - \mathbf{x})}_{\mathbf{w}' \in W},$$

shows that the elements of $V + W$ may have several representations as sums $\mathbf{v} + \mathbf{w}$ $(\mathbf{v} \in V, \ \mathbf{w} \in W)$. Uniqueness of decompositions requires $V \cap W = \{0\}$. Conversely, when this condition is satisfied,

$$\mathbf{v} + \mathbf{w} = \mathbf{v}' + \mathbf{w}' \quad (\mathbf{v}, \mathbf{v}' \in V; \ \mathbf{w}, \mathbf{w}' \in W)$$

implies

$$\mathbf{v} - \mathbf{v}' \in V, \quad \mathbf{v} - \mathbf{v}' = \mathbf{w}' - \mathbf{w} \in W,$$

hence $\mathbf{v} - \mathbf{v}' \in V \cap W = \{0\}$, $\mathbf{v} = \mathbf{v}'$ and $\mathbf{w}' = \mathbf{w}$, namely uniqueness.

Example. Let E be the usual 3-dimensional space, V the horizontal plane defined by $z = 0$, and W the vertical plane defined by $y = 0$. Then each vector of $\mathbf{R}^3$ has a decomposition

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ z \end{pmatrix},$$

with components in V (resp. W). This proves $V + W = \mathbf{R}^3$. But the same vector may be decomposed in several ways

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} x + a \\ y \\ 0 \end{pmatrix} + \begin{pmatrix} -a \\ 0 \\ z \end{pmatrix}$$

(a any scalar), with components in V resp. W. Different decompositions differ by a choice of element

$$\begin{pmatrix} a \\ 0 \\ 0 \end{pmatrix} \in V \cap W\colon\ y = 0 \text{ and } z = 0.$$

Proposition. *The following properties concerning two subspaces V and W of a vector space E, are equivalent:*

(*i*) *If A is a basis of V and B a basis of W, then $A \cup B$ is a basis of $V + W$*

(*ii*) *If A (resp. B) is an independent subset of V (resp. W) then $A \cup B$ is an independent subset of $V + W$*

(*iii*) $\mathbf{v} + \mathbf{w} = 0$ $(\mathbf{v} \in V,\ \mathbf{w} \in W)$ *implies* $\mathbf{v} = \mathbf{w} = 0$

(*iv*) *Each $\mathbf{x} \in V + W$ has a* unique *decomposition* $\mathbf{x} = \mathbf{v} + \mathbf{w}$ $(\mathbf{v} \in V,\ \mathbf{w} \in W)$

(*v*) $V \cap W = \{0\}$.

PROOF. Since the proof follows easily from the preliminary considerations, we leave the assiduous reader write it in detail. ∎

Notice that when the subspaces V and W are finite dimensional, a condition equivalent to (*i*) is

$$\dim(V + W) = \dim V + \dim W.$$

Definition. *Two subspaces V and W of a vector space E are* independent *when $V \cap W = \{0\}$, hence all the conditions of the preceding proposition are satisfied.*

A good way of constructing examples of independent subspaces of E is as follows. Take a basis $(\mathbf{e}_i)_{i \in I}$ of E. Then for disjoint subsets K, $L \subset I$, the subspaces generated by the corresponding basis elements

$$V = \mathcal{L}(\mathbf{e}_i : i \in K), \quad W = \mathcal{L}(\mathbf{e}_i : i \in L),$$

are independent. When $I = K \cup L$, then the corresponding subspaces have a sum equal to E.

5.2.2 Supplementary Subspaces

A basis of the space $\Pi_{<n}$ of polynomials of degrees less than n is $1, x, \ldots, x^{n-1}$. The monomials $x^n, x^{n+1}, \ldots$ generate an independent subspace W_n (Sec. 2.3.1) of the space Π of all polynomials, such that $\Pi_{<n} + W_n = \Pi$.

Quite generally, let V be a subspace of a vector space E. It follows from the *mathematical credo* (Sec. 2.3.2) that there is a basis $(\mathrm{e}_i)_{i\in J}$ of V, and that we can complete it into a basis $(\mathrm{e}_i)_{i\in I}$ of E. The elements $(\mathrm{e}_i)_{i\in I-J}$ make up a basis of a subspace W of E which is independent of V and such that $E = V+W$.

Definition. *Two subspaces V, W of a vector space E are* supplementary subspaces, *or* supplement of each other, *when they are independent and generate E, namely when*

$$V \cap W = \{0\} \quad \textit{and} \quad V + W = E.$$

Example. Let $f = p/q$ be a rational function, written as a quotient of two polynomials p and q. The Euclidean division of p by q (according to decreasing powers of the variable x) leads to two uniquely determined polynomials m and r satisfying

$$p = mq + r \quad (\deg r < \deg q).$$

Hence

$$\frac{p}{q} = m + \frac{r}{q} \quad (\deg r < \deg q).$$

This shows that the subset—a subspace—of rational functions $g = \frac{r}{q}$ (with $\deg r < \deg q$) is a supplement of the subspace of polynomials. If we work with real (or complex) coefficients, this supplement consists of the rational functions g such that $g(x) \to 0$ when $x \to \infty$.

As we have seen, it follows from the mathematical credo that *each subspace V of a vector space E has a supplement W, and there is a surjective linear map with domain E and kernel V, namely*

$$\begin{aligned} f\,:\,E &\longrightarrow W \\ x = y + z &\longmapsto z \quad (y \in V,\, z \in W). \end{aligned}$$

Proposition. *Let $f : E \to F$ be a linear map, and W a subspace of E. The restriction of f to W is injective precisely when W and* $\ker f$ *are independent. The restriction $f|_W$ to any supplement W of* $\ker f$ *furnishes an isomorphism $W \xrightarrow{\sim} \operatorname{im} f$.*

PROOF. The restriction of f to a subspace W is injective when

$$(\ker f) \cap W = \ker f|_W = \{0\},$$

namely when W is independent from $\ker f$. On the other hand

$$\begin{aligned} \operatorname{im} f|_W = f(W) &= f(V+W) && (\text{if } V \subset \ker f) \\ &= f(E) = \operatorname{im} f && (\text{if } V + W = E), \end{aligned}$$

so that the proposition follows. ■

Corollary 1. *Let E be a vector space, V a subspace of E. Then two supplements of V in E are isomorphic.*

PROOF. As we have observed, there is a linear map $f : E \to F$ having V as kernel. The proposition shows that if W is a supplement of V, then the restriction of f furnishes an isomorphism $W \xrightarrow{\sim} \operatorname{im} f$. If W' is also a supplement of V, we have

$$\begin{array}{ccc} E \supset W & \searrow^{\sim} & \\ & & \operatorname{im} f, \\ E \supset W' & \nearrow^{\sim} & \end{array}$$

so that W and W' are isomorphic. ∎

Corollary 2. *Two linear maps $f : E \to F$, $f' : E \to F'$ having the same kernel, have isomorphic images.*

PROOF. Let us choose a supplement W to the common kernel V of f and f'. Then by the proposition, both images are isomorphic to W:

$$\begin{array}{ccc} & \nearrow^{\sim} & \operatorname{im} f \subset F \\ E \supset W & & \\ & \searrow^{\sim} & \operatorname{im} f' \subset F', \end{array}$$

hence are isomorphic. ∎

Comments. Recall that the rank-nullity theorem (Sec. 4.3.1) was just proved by constructing a supplement of $\ker f$. In fact, the above proposition constitutes a generalization (without finiteness assumption) of the rank-nullity theorem: The decomposition $E = \ker f + W$, $\ker f \cap W = \{0\}$ for a finite-dimensional vector space E, leads to

$$\dim E = \dim \ker f + \underbrace{\dim W}_{=\dim \operatorname{im} f},$$

in view of the isomorphism $W \xrightarrow{\sim} \operatorname{im} f$. In Sec. 7.4.2, we shall show how a natural supplement W of $\ker f$ can be obtained with an inner product, hence a natural isomorphism between the row and column spaces of (real, or complex) matrices. The above second corollary is to be compared to the main property of row operations:

A and $A' = EA \sim A$ have the same kernel hence isomorphic images.

Definition. *The* codimension *of a subspace V of a vector space E is the dimension of any supplement W of V in E.*

In general, this codimension may be infinite. But even if V is infinite dimensional, its codimension may be finite. For example, if $\varphi : E \to \mathbf{R}$ is a nonzero linear form, its kernel has codimension 1. Indeed, any supplement to $\ker \varphi$ is isomorphic to the one-dimensional space $\mathbf{R}$.

When V is a subspace of a finite-dimensional vector space E, then for any supplement W of V

$$\dim V + \dim W = \dim E,$$

hence the codimension of V is $\dim E - \dim V$, whence the terminology.

5.2.3 Direct Sum of Two Subspaces

Another useful equivalent formulation of independence for two subspaces consists in introducing the *direct-sum* vector space $V \oplus W$ having for elements the pairs $(\mathbf{v}, \mathbf{w})$ where $\mathbf{v} \in V$ and $\mathbf{w} \in W$. Addition of such pairs is defined componentwise, multiplication by scalars similarly. Then the addition of components defines a natural *sum map*

$$\begin{aligned} \Sigma : V \oplus W &\longrightarrow E \\ (\mathbf{v}, \mathbf{w}) &\longmapsto \mathbf{v} + \mathbf{w}. \end{aligned}$$

This map is linear since

$$a(\mathbf{v}, \mathbf{w}) + (\mathbf{v}', \mathbf{w}') = (a\mathbf{v} + \mathbf{v}', a\mathbf{w} + \mathbf{w}')$$

(by definition of the composition laws in $V \oplus W$) maps to

$$(a\mathbf{v} + \mathbf{v}') + (a\mathbf{w} + \mathbf{w}') = a(\mathbf{v} + \mathbf{w}) + (\mathbf{v}' + \mathbf{w}')$$

(by the axiomatic properties valid in the vector space E), and this is

$$a\Sigma(\mathbf{v}, \mathbf{w}) + \Sigma(\mathbf{v}', \mathbf{w}'),$$

as required for Σ to be linear. The image of Σ is by definition the subspace $V + W$. By definition, V and W are independent when Σ is an isomorphism

$$\begin{aligned} V \oplus W &\xrightarrow{\sim} V + W \subset E \\ (\mathbf{v}, \mathbf{w}) &\longmapsto \mathbf{v} + \mathbf{w}. \end{aligned}$$

In this case, it is convenient to identify the direct sum with its image in E, and write $V \oplus W \subset E$. We may even say that a sum $V + W$ of subspaces *is a direct sum* when it is isomorphic to $V \oplus W$, namely when $V \cap W = \{0\}$. Hence, E *is the direct sum* of V and W when they are supplementary subspaces

$$E = V \oplus W \quad \Longleftrightarrow \quad V + W = E \text{ and } V \cap W = \{0\}.$$

When V and W are finite dimensional, it is obvious that $V \oplus W$ is also finite dimensional, and

$$\dim(V \oplus W) = \dim V + \dim W.$$

Proposition. *Let V and W be two finite-dimensional subspaces of E. Then $V + W$ is also finite dimensional, and*

$$\dim(V + W) = \dim V + \dim W - \dim(V \cap W).$$

PROOF. The sum map $\Sigma : V \oplus W \to E$ has image $V + W$. Its kernel consists of the pairs $(\mathbf{v}, \mathbf{w})$ with $\mathbf{v} + \mathbf{w} = 0$, $\mathbf{w} = -\mathbf{v} \in V \cap W$, hence is isomorphic to $V \cap W$ via

$$V \cap W \xrightarrow{\sim} \ker \Sigma, \quad \mathbf{v} \longmapsto (\mathbf{v}, -\mathbf{v}).$$

By the rank-nullity theorem, we have

$$\begin{aligned} \dim(V \oplus W) &= \dim \operatorname{im} \Sigma + \dim \ker \Sigma, \\ \dim V + \dim W &= \dim(V + W) + \dim(V \cap W). \end{aligned}$$

The announced formula follows. ∎

Application. Let us show how to construct a basis of the intersection of two subspaces V, W of $\mathbf{R}^n$ from given bases $(\mathbf{v}_i)_{1 \leqslant i \leqslant p}$ of V and $(\mathbf{w}_j)_{1 \leqslant j \leqslant q}$ of W. Consider the matrix

$$A = \big(\mathbf{v}_1, \ldots, \mathbf{v}_p \mid \mathbf{w}_1, \ldots, \mathbf{w}_q\big),$$

of size $n \times (p + q)$. Since the columns of this matrix generate the subspace

$$\mathcal{L}(\mathbf{v}_1, \ldots, \mathbf{v}_p, \mathbf{w}_1, \ldots, \mathbf{w}_q) = V + W,$$

the rank of A is $r = \dim(V + W)$. By row operations, we can bring A into a row-reduced echelon form $U = EA$. The isomorphism produced by left multiplication by E in $\mathbf{R}^n$ moves the spaces under consideration into a simple position, in which the problem is more easily solved. Since row operations preserve linear relations of columns, the first p columns of U are independent, hence are $\mathbf{e}_1, \ldots, \mathbf{e}_p$, and U has the form

$$\begin{aligned} U = EA &= \big(E\mathbf{v}_1, \ldots, E\mathbf{v}_p \mid E\mathbf{w}_1, \ldots, E\mathbf{w}_q\big) \\ &= \big(\mathbf{e}_1, \ldots, \mathbf{e}_p \mid \mathbf{u}_1, \ldots, \mathbf{u}_q\big), \end{aligned}$$

where the s nonpivot columns occur in the second block. For the same reason, the last q columns of U are also independent, while the pivot columns in this block are $\mathbf{e}_{p+1}, \ldots, \mathbf{e}_r$. As we see, the isomorphism produced by multiplication by E replaces V by $EV = \mathbf{R}^p$ (with its canonical basis), and W by EW (with a basis containing $r - p$ canonical elements). The equality

$$\begin{aligned} \dim V + \dim W &= \dim(V + W) + \dim(V \cap W) \\ p + q &= r + s, \end{aligned}$$

shows that

$$s = \dim(V \cap W)$$

is the number of nonpivot columns of U (or A). *I claim that the truncations* $\mathbf{u}'_j \in \mathbf{R}^p = EV$ *of the nonpivot columns* $\mathbf{u}_j$ *give a basis of the intersection of* EV *and* EW. Each $\mathbf{u}'_j$ is obtained from the corresponding $\mathbf{u}_j$ by forgetting its last components

$$\mathbf{R}^p \ni \mathbf{u}'_j = \mathbf{u}_j - \sum_{p+1 \leqslant k \leqslant r} u_{kj} \mathbf{e}_k \in EW,$$

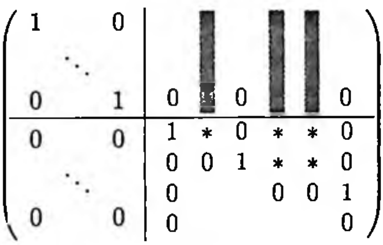
hence lies in $\mathbf{R}^p \cap EW$. Since these linear combinations, together with the pivot columns $\mathbf{e}_{p+1}, \ldots, \mathbf{e}_r$ generate EW, they form a basis of this space. This proves that the $\mathbf{u}'_j$ (corresponding to nonpivot columns) are independent and belong to the s-dimensional space $EV \cap EW$: They form a basis of this space. By definition, the components of the $\mathbf{u}'_j$ are the first r components of the nonpivot columns, so that we can write

$$\mathbf{u}'_j = \sum_{1 \leqslant k \leqslant r} u_{kj} \mathbf{e}_k = \sum_{1 \leqslant k \leqslant r} u_{kj} E\mathbf{v}_k.$$

Using the inverse isomorphism E^{-1}, we find that *the vectors*

$$\mathbf{v}'_j = \sum_{1 \leqslant k \leqslant r} u_{kj} \mathbf{v}_k \quad (j \text{ nonpivot column}),$$

form a basis of the intersection of V and W. In the following picture, three nonpivot columns lead to a basis of the intersection: The coefficients of the $\mathbf{v}_i$'s are found in the shaded rectangles.



5.2.4 Independent Subspaces (General Case)

The preceding definition of independence for two subspaces may be generalized to arbitrary families as follows.

Definition. *A family $(V_i)_{i \in I}$ of subspaces $V_i \subset E$ is called* independent, *when a finite sum $\sum \mathbf{v}_i$ of elements $\mathbf{v}_i \in V_i$ can vanish only if $\mathbf{v}_i = 0$ for all i.*

As we shall mainly be concerned with finite families of subspaces, we shall restrict the index set to be finite, or $\mathbf{N}$. This is only a notational simplification. (Recall however the infinite-dimensional examples in Sec. 2.3.3, where the relevance of more general index sets appears.)

Proposition. *Let $(V_i)_{i \geqslant 0}$ be an independent family of subspaces of E. Choose subsets $S_i \subset V_i$ $(i \geqslant 0)$. If each S_i is linearly independent, then the union $\bigcup_{i \geqslant 0} S_i \subset E$ is also linearly independent.*

PROOF. Let $S_0 = \{\mathbf{e}_i : i \in I_0\} \subset V_0$, $S_1 = \{\varepsilon_j : j \in I_1\} \subset V_1\}, \ldots$ be linearly independent subsets. I claim that the union of these sets is also linearly independent in E. Consider any linear relation

$$\underbrace{\textstyle\sum_i a_i \mathbf{e}_i}_{\mathbf{v}_0 \in V_0} + \underbrace{\textstyle\sum_j b_j \varepsilon_j}_{\mathbf{v}_1 \in V_1} + \cdots = 0$$

(finitely many components, each one containing at most finitely many nonzero elements). By definition of independence of the subspaces V_i

$$\mathbf{v}_0 = 0, \quad \mathbf{v}_1 = 0, \quad \ldots$$

By linear independence of $(\mathbf{e}_i)$ in V_0, the first equality $\mathbf{v}_0 = \sum_i a_i \mathbf{e}_i = 0$ implies that all a_i are zero. By linear independence of (ε_j) in V_1, the second equality $\mathbf{v}_2 = \sum_j b_j \varepsilon_j = 0$ implies similarly that all b_j are zero, and so on. ■

Corollary. *Let $(V_i)_{i \geqslant 0}$ be an independent family of nonzero vector subspaces in a finite-dimensional space E. Then this family is finite and*

$$\textstyle\sum_i \dim V_i \leqslant n = \dim E.$$

PROOF. The union of bases $B_i \subset V_i$ is a linearly independent subset of E, hence has at most n elements. ■

The dimension of the sum of two finitely generated subspaces (Sec. 5.2.3)

$$\dim V + \dim W = \dim(V + W) + \dim(V \cap W),$$

can also be derived from the preceding proposition. Take a supplement V_1 of $V \cap W$ in V, as well as a supplement W_1 of $V \cap W$ in W. Then the sums

$$V = (V \cap W) + V_1, \quad W = (V \cap W) + W_1$$

are direct sums, so that

$$\dim V = \dim(V \cap W) + \dim V_1, \quad \dim W = \dim(V \cap W) + \dim W_1.$$

Hence

$$\dim V + \dim W = \dim V_1 + 2\dim(V \cap W) + \dim W_1.$$

Since V_1, $V \cap W$, and W_1 are independent with sum $V + W$, we have

$$\dim V_1 + \dim(V \cap W) + \dim W_1 = \dim(V + W),$$

whence the result.

5.2.5 Finite Direct Sums of Vector Spaces

Let $(E_i)_{1 \leqslant i \leqslant \ell}$ be a finite family of vector spaces. By definition, the *direct sum* of these spaces consists of the families $(\mathbf{v}_1, \ldots, \mathbf{v}_\ell)$, where $\mathbf{v}_i \in E_i$ for all i. There is no basic difference with the product, although we write elements of products in column, to conform with the convention concerning $\mathbf{R}^n$ as consisting of column n-tuples. For example, the direct sum of n copies of $\mathbf{R}$ is $\mathbf{R}_n$. The direct sum is a vector space with respect to componentwise addition and multiplication by scalars

$$\begin{aligned}
(\mathbf{v}_1, \ldots, \mathbf{v}_\ell) + (\mathbf{v}'_1, \ldots, \mathbf{v}'_\ell) &= (\mathbf{v}_1 + \mathbf{v}'_1, \ldots, \mathbf{v}_\ell + \mathbf{v}'_\ell), \\
a(\mathbf{v}_1, \ldots, \mathbf{v}_\ell) &= (a\mathbf{v}_1, \ldots, a\mathbf{v}_\ell) \quad (a \in \mathbf{R}),
\end{aligned}$$

and is denoted by $E_1 \oplus \cdots \oplus E_\ell$ or $\bigoplus_{1\leqslant i\leqslant \ell} E_i$. From bases of the components E_i, we easily deduce a basis of the direct sum. If all E_i's are finite dimensional, we infer

$$\dim(E_1 \oplus \cdots \oplus E_\ell) = \dim E_1 + \cdots + \dim E_\ell.$$

If the $E_i = V_i \subset E$ are subspaces of a fixed space E, the addition in E furnishes a linear map

$$\begin{array}{rcl} \Sigma : V_1 \oplus \cdots \oplus V_\ell & \longrightarrow & E \\ (\mathbf{v}_1, \ldots, \mathbf{v}_\ell) & \longmapsto & \mathbf{v}_1 + \cdots + \mathbf{v}_\ell. \end{array}$$

Its image is by definition the sum of the subspaces V_i in E, and

$$\begin{aligned} V_1 + \cdots + V_\ell &= \sum_{1\leqslant i\leqslant \ell} V_i = \{\mathbf{v}_1 + \cdots + \mathbf{v}_\ell : \mathbf{v}_i \in V_i \text{ for all } i\}, \\ \dim(V_1 + \cdots + V_\ell) &\leqslant \dim(V_1 \oplus \cdots \oplus V_\ell) = \dim V_1 + \cdots + \dim V_\ell, \end{aligned}$$

by the rank-nullity theorem. The kernel of Σ corresponds to the dependence relations between elements of the V_i's.

Proposition 1. *A family $(V_i)_{1\leqslant i\leqslant \ell}$ of subspaces of E is* independent*when the linear map $(\mathbf{v}_1, \ldots, \mathbf{v}_\ell) \mapsto \mathbf{v}_1 + \cdots + \mathbf{v}_\ell$ given by addition is injective, hence an isomorphism with its image*

$$V_1 \oplus \cdots \oplus V_\ell \xrightarrow{\sim} V_1 + \cdots + V_\ell \subset E. \qquad \blacksquare$$

If all the subspaces V_i are finite dimensional, they are independent precisely when

$$\dim(V_1 + \cdots + V_\ell) = \dim V_1 + \cdots + \dim V_\ell.$$

When this is the case, the corresponding sum is also called *direct* and identified with $\bigoplus_{1\leqslant i\leqslant \ell} V_i$. The important point is

> *any element* $\mathbf{x}$ *in a direct sum has a* unique decomposition
> $\mathbf{x} = \sum_i \mathbf{v}_i$ *with* $\mathbf{v}_i \in V_i$ *for all* i.

Remark. The independence of subspaces $V_1, \ldots, V_n$ implies that the intersections of any two different ones is reduced to $\{0\}$: For $i \neq j$

$$\mathbf{x} \in V_i \cap V_j \quad \Longrightarrow \quad \underset{\in V_i}{(\mathbf{x})} + \underset{\in V_j}{(-\mathbf{x})} = 0 \quad \Longrightarrow \quad \mathbf{x} = 0.$$

But the conditions $V_i \cap V_j = \{0\}$ for all $i \neq j$ are not sufficient for global independence. For example, three different homogeneous lines $L_i \subset \mathbf{R}^2$ intersect in the origin only. But they cannot be independent since we can find nontrivial relations $\vec{v}_1 + \vec{v}_2 + \vec{v}_3 = 0$ $(0 \neq \vec{v}_i \in L_i)$ in the 2-dimensional space $\mathbf{R}^2$. The same happens with four homogeneous lines in $\mathbf{R}^3$. Here is a criterion for independence of finite families of subspaces of a vector space.

Proposition 2. *A family $(V_i)_{1\leqslant i\leqslant \ell}$ of subspaces of E is independent when*

$$V_i \cap \sum_{j\neq i} V_j = \{0\} \quad (1\leqslant i\leqslant \ell).$$

PROOF. Assume that the family $(V_i)_{1\leqslant i\leqslant \ell}$ is independent. Fix an index i and take $\mathbf{v} = \mathbf{v}_i = \sum_{j\neq i}\mathbf{v}_j \in V_i \cap \sum_{j\neq i} V_j$. Hence

$$\mathbf{v}_i - \sum\nolimits_{j\neq i}\mathbf{v}_j = 0$$

implies $\mathbf{v}_i = 0$ since the subspaces are independent: Hence $V_i \cap \sum_{j\neq i} V_j = \{0\}$. Conversely, assume $V_i \cap \sum_{j\neq i} V_j = \{0\}$, and consider a relation $\sum_j \mathbf{v}_j = 0$ where $\mathbf{v}_j \in V_j$ for all j. We may rewrite it as

$$\mathbf{v}_i + \sum\nolimits_{j\neq i}\mathbf{v}_j = 0,$$

or as

$$\mathbf{v}_i = -\sum\nolimits_{j\neq i}\mathbf{v}_j \in V_i \cap \sum\nolimits_{j\neq i} V_j = \{0\} : \quad \mathbf{v}_i = 0.$$

Since $V_i \cap \sum_{j\neq i} V_j = \{0\}$ for all i, we also deduce $\mathbf{v}_i = 0$ for all i, hence the independence of the subspaces V_i's. ■

We recover the initial definition for two subspaces V_1, $V_2 \subset E$: They are independent precisely when $V_1 \cap V_2 = \{0\}$.

5.3 Projectors

5.3.1 An Example and General Definition

Let us start by an example in the usual space $\mathbf{R}^3$. Consider the *vertical projection P* on the plane of equation $z = ax - by$. This is a linear map and we can determine its matrix in the canonical basis $(\vec{e}_i)$ of $\mathbf{R}^3$. By definition, a vertical projection preserves the first two components of any vector. In particular, the image of the first basis vector has the form $P\vec{e}_1 = {}^t(1,0,?)$. We have to compute the third component. Since this image is in the plane, we have

$$? = a1 - b0 = a,$$

and $P\vec{e}_1 = {}^t(1,0,a)$. This is the first column in the matrix description of P. The image of the second basis vector is determined in a similar way

$$P\vec{e}_2 = {}^t(0,1,-b).$$

Finally, the third basis element $\vec{e}_3$ is vertical, so that $P\vec{e}_3 = 0$ by definition. Here is the matrix of P

$$M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ a & -b & 0 \end{pmatrix}.$$

By definition of a projection, P is the identity on its image: $P^2 = P$. One can check $M^2 = M$.

Definition. Let E be a vector space. We call projector *in E any linear map $P : E \to E$ such that $P^2 = P$.*

When $\dim E \geqslant 2$, the quadratic equation $P^2 = P$ has infinitely many solutions P which are linear maps $E \to E$. (Recall that the quadratic equation $X^2 = -I$ also has infinitely many solutions X which are 2×2 matrices.)

Example. Let E be the space of functions $f : \mathbf{R} \to \mathbf{R}$. For any such function, define f_s by $f_s(x) = f(-x)$, and consider the linear map: $E \longrightarrow E$

$$f \longmapsto Pf = \tfrac{1}{2}(f + f_s).$$

The image $g = Pf$ is the average of f and its symmetric: It is an *even* function. Recall that even functions g are characterized by

$$g(-x) = g(x) \quad (x \in \mathbf{R}).$$

For an even function g, $Pg = g$, and hence $P^2 f = Pg = g = Pf$. This proves $P^2 = P$ so that P is a projector. In fact, $f = Pf$ precisely when f is even, so that the image of f is the subspace V of even functions, on which P acts by the identity. On the other hand, if h is an *odd* function, namely

$$h(-x) = -h(x) \quad (x \in \mathbf{R}),$$

then $Ph = 0$: The kernel of P is the subspace of odd functions. Any function f is the sum of an even one $g = Pf$ and an odd one $h = Qf = \frac{1}{2}(f - f_s)$ and

$$Pf = P(g + h) = Pg + Ph = Pg = g.$$

5.3.2 Geometrical Meaning of $P^2 = P$

Let us discover the geometrical meaning of the algebraic condition $P^2 = P$ for a linear map $P : E \to E$ of a vector space in itself. The two subspaces

$$V = P(E) = \operatorname{im} P, \quad W = \ker P$$

will play a leading role. If $\mathbf{x} \in V$, we can write $\mathbf{x} = P\mathbf{y}$ for some $\mathbf{y} \in E$ and

$$P\mathbf{x} = P(P\mathbf{y}) = P^2\mathbf{y} \overset{!}{=} P\mathbf{y} = \mathbf{x}.$$

This shows that P acts by the identity in the subspace V: $P|_V = id|_V$. On the other hand, P acts trivially in its kernel: $P|_W = O|_W$. Grouping these two facts, we deduce that if $\mathbf{x}$ has the form $\mathbf{y} + \mathbf{z}$ with $\mathbf{y} \in V$ and $\mathbf{z} \in W$, then $P\mathbf{x} = \mathbf{y}$ is determined by additivity

$$\begin{array}{rccccl} \mathbf{x} = & \mathbf{y} & + & \mathbf{z} & & \\ P & \downarrow & & \downarrow & & \\ & \mathbf{y} & + & 0 & = \mathbf{y}. \end{array}$$

It turns out that *any element* $\mathbf{x} \in E$ *is of the preceding form* $\mathbf{y} + \mathbf{z}$. Here is the reason. For any $\mathbf{x} \in E$, define $\mathbf{y} = P\mathbf{x} \in V$ and $\mathbf{z} = \mathbf{x} - P\mathbf{x}$, in order to have $\mathbf{x} = \mathbf{y} + \mathbf{z}$. Then we have

$$P\mathbf{z} = P(\mathbf{x} - P\mathbf{x}) = P\mathbf{x} - P^2\mathbf{x} = P\mathbf{x} - P\mathbf{x} = 0 : \quad \mathbf{z} \in \ker P = W.$$

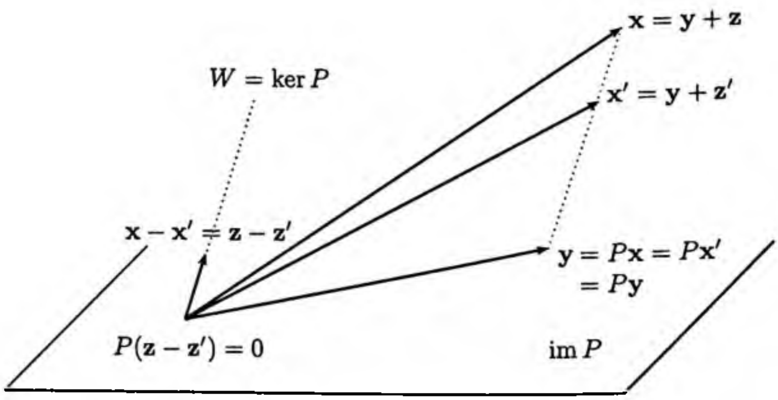
This decomposition of $\mathbf{x} \in E$

$$\mathbf{x} = \mathbf{y} + \mathbf{z} \quad (\mathbf{y} \in V, \mathbf{z} \in W)$$

is the only one since $V \cap W = \{0\}$. In this decomposition, the action of P is fully described by

$$\begin{array}{ccccc} \mathbf{x} & = & \mathbf{y} & + & \mathbf{z} \\ \downarrow & & \downarrow & & \downarrow \\ P\mathbf{x} & = & \mathbf{y} & + & 0. \end{array}$$

Let us say that P projects E onto V, in the direction parallel to W. Here is a picture of a plane projector in the usual 3-dimensional space.



Recall (Sec. 5.2.3): A space E is the *direct sum* of two supplementary subspaces V and W when each element $\mathbf{x} \in E$ can be written in a *unique* way

$$\mathbf{x} = \mathbf{y} + \mathbf{z} \quad (\mathbf{y} \in V, \mathbf{z} \in W),$$

namely when

$$E = V + W \quad \text{and} \quad V \cap W = \{0\}.$$

Hence we have proved the following proposition.

Proposition. *Let $P : E \to E$ be a projector: $P^2 = P$. Then E is a direct sum of $V = \operatorname{im} P$ and $W = \ker P$, with $P|_V = id_V$ and $P|_W = O|_W$.* ■

There is a 1-1 correspondence between the projectors P in E and the ordered pairs (V, W) consisting of two supplementary subspaces of E. It is given by

$$P \longmapsto (\operatorname{im} P, \ker P).$$

When P is a projector in E, define $Q = I_E - P$. Since

$$\begin{aligned} Q^2 &= I_E - I_E P - P I_E + P^2 \\ &= I_E - 2P + P \ = \ I_E - P \ = \ Q, \end{aligned}$$

Q is also a projector. By definition, $Q\mathbf{x} = \mathbf{x} - P\mathbf{x}$, so that

$$\begin{aligned} Q\mathbf{x} = \mathbf{x} \quad &\Longleftrightarrow \quad P\mathbf{x} = 0 \\ Q\mathbf{x} = 0 \quad &\Longleftrightarrow \quad P\mathbf{x} = \mathbf{x}, \end{aligned}$$

and

$$\operatorname{im} Q = \ker P, \quad \ker P = \operatorname{im} Q.$$

The relation $P+Q = I$ shows that P and Q play symmetric roles: It is sufficient to notice that the kernel of P is the image of Q to be able to deduce that the image of P is the kernel of Q (interchange the roles of P and Q). The symmetry in P and Q (or V and W) in the diagrams

$$\begin{array}{ccccccc} E & \xrightarrow{\sim} & V \oplus W & \qquad & V \oplus W & \xrightarrow{\sim} & E \\ \mathbf{x} & \longmapsto & (P\mathbf{x}, Q\mathbf{x}) & & (\mathbf{y}, \mathbf{z}) & \longmapsto & \mathbf{y} + \mathbf{z} \end{array}$$

is an invitation to generalize the preceding considerations to finite families of projectors.

Proposition. *Let $P_1, \ldots, P_\ell$ be a family of projectors in E such that*

$$P_i P_j = P_j P_i = 0 \textit{ for all } i \neq j.$$

Then $P = P_1 + \cdots + P_\ell$ is a projector. The family of subspaces $V_i = \operatorname{im} P_i$ is independent and P is a projector onto the direct sum $\sum_i V_i$.

PROOF. Consider the algebraic identity

$$(P_1 + \cdots + P_\ell)^2 = \textstyle\sum_{i,j} P_i P_j = P_1^2 + \cdots + P_\ell^2 + \sum_{i \neq j} P_i P_j.$$

With our assumptions $P_i P_j = \delta_{ij} P_i$, we get

$$(P_1 + \cdots + P_\ell)^2 = P_1^2 + \cdots + P_\ell^2 = P_1 + \cdots + P_\ell,$$

whence the result. ∎

When $P_1, \ldots, P_\ell$ is a family of projectors in E such that

$$P_i P_j = P_j P_i = 0 \quad (i \neq j),$$

we may replace E by $\operatorname{im} P = \operatorname{im}(P_1 + \cdots + P_\ell)$ and thus assume

$$P_1 + \cdots + P_\ell = I_E.$$

In this case, we have an isomorphism

$$\begin{array}{rcl} E & \xrightarrow{\sim} & V_1 \oplus \cdots \oplus V_\ell \\ \mathbf{x} & \longmapsto & (P_1\mathbf{x}, \ldots, P_\ell\mathbf{x}), \end{array}$$

with inverse given by the sum map

$$\begin{array}{rcl} V_1 \oplus \cdots \oplus V_\ell & \xrightarrow{\sim} & E \\ (\mathbf{v}_1, \ldots, \mathbf{v}_\ell) & \longmapsto & \mathbf{v}_1 + \cdots + \mathbf{v}_\ell. \end{array}$$

5.3.3 Tricks of the Trade

Here are a few *algebraic properties of projectors*. If P and Q are two projectors, then

$$P + Q \text{ projector} \iff PQ + QP = 0.$$

Indeed $P^2 = P$ and $Q^2 = Q$ by assumption, so that

$$P + Q = (P + Q)^2 = P^2 + PQ + QP + Q^2 \iff PQ + QP = 0.$$

More can be said.

Proposition 1. *If P and Q are two projectors, then*

$$P + Q \textit{ projector} \iff PQ = QP = 0.$$

PROOF. As we have just seen, $P+Q$ is a projector precisely when $PQ+QP = 0$. If $PQ = QP = 0$, this condition is verified. Conversely, if $PQ + QP = 0$, we can multiply this identity on the left and on the right by Q, obtaining

$$\begin{cases} QPQ + Q^2P &= QPQ + QP = 0 \\ PQ^2 + QPQ &= PQ + QPQ = 0. \end{cases}$$

This shows $QP = -QPQ = PQ$, and $2PQ = 0$ (since $PQ + QP = 0$). Finally, we see that $QP = PQ = 0$ as expected. ∎

Proposition 2. *If P and Q are two projectors, then*

$$P - Q \textit{ projector} \iff PQ = QP = Q.$$

PROOF. As we have seen, $P - Q$ is a projector exactly when $I - (P - Q) = (I - P) + Q$ is a projector. Using the preceding proposition, we see that this is the case precisely when

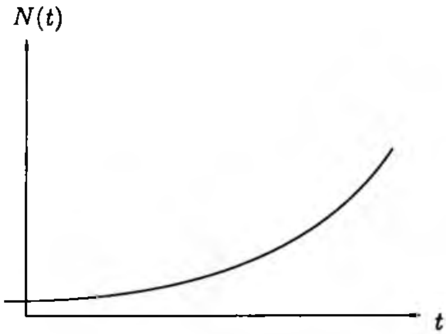
$$(I - P)Q = Q(I - P) = 0,$$

namely when $Q = PQ$ and $Q = QP$. ∎

5.4 Appendix

5.4.1 Pyramid of Ages

Most countries spend a lot of money in the statistical study of their population and its growth. A first approach consists in evaluating the population as a positive integer. If this integer is large, its evolution as a function of time is best described by a Cartesian graph where time, as well as population are treated as real numbers. Newspapers mention *exponential growth* in this context (even when this is an understatement: With an exponential growth, the doubling time is constant; but the doubling time for the world population—close to 1000 years before Christ—has diminished dramatically to about 30 years around 1980, and was estimated to 47 years in 1997.)

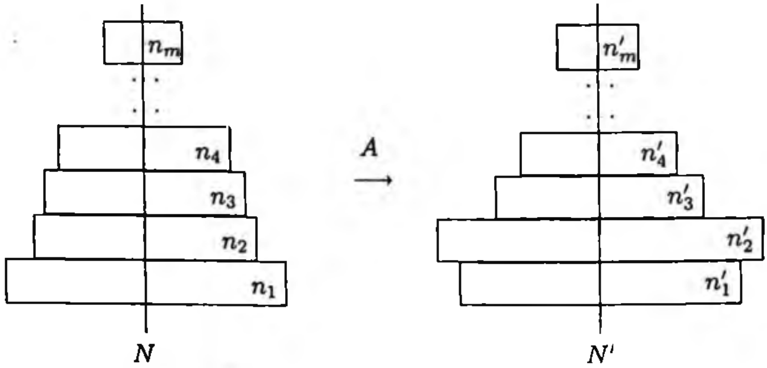


In a more sophisticated approach, statistics considers the breaking down of the population into age groups. Instead of the total number $N(t)$ of population at time t, it is more informative to list the partial numbers $n_i(t)$ in different age groups $(1 \leqslant i \leqslant m)$

$$N(t) = n_1(t) + n_2(t) + \cdots + n_m(t).$$

This data is an m-tuple, hence a generalized vector. The partition between young, productive adults, and senior citizen is quite common. But statistics offices use a finer partition with 20 groups of 5 years each. (They even separate men and women, single, married, divorced, etc. thus achieving a quite large matrix of data: Here, we only consider the principle, explained in a simplified situation.) Hence we may say that such a model involves the 20-dimensional vector space $\mathbf{R}^{20}$.

Let us now approach a modelling of the evolution of such a population. Assume that the evolution from one generation to the next is such that the new born are produced by the different age groups



$$n'_1 = f_1 n_1 + f_2 n_2 + \cdots + f_m n_m,$$

with specific fertilities (say $f_1 = f_2 = 0$, $f_m = 0$). The next age groups n'_k ($k \geqslant 2$) come from the preceding ones n_{k-1}

$$n'_k = p_k n_{k-1} \quad (2 \leqslant k \leqslant m),$$

with certain probabilities of survival p_j. (This model ignores both emigration and immigration.) These relations can be gathered in a matrix product

$$\begin{pmatrix} n'_1 \\ n'_2 \\ \vdots \\ \vdots \\ n'_m \end{pmatrix} = \begin{pmatrix} 0 & f_2 & f_3 & \dots & f_m \\ p_2 & 0 & 0 & \dots & 0 \\ 0 & p_3 & 0 & \dots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & & p_m & 0 \end{pmatrix} \begin{pmatrix} n_1 \\ n_2 \\ \vdots \\ \vdots \\ n_m \end{pmatrix},$$

or in short

$$\mathbf{n}' = L\mathbf{n}.$$

The particular matrices L that appear in this model are called *Leslie matrices*.

Under the assumption that the fertilities and survival probabilities remain constant, the evolution after k generations is described by the kth power of the corresponding Leslie matrix

$$\mathbf{n}^{(k)} = L^k \mathbf{n} \quad (k \geqslant 1).$$

It is interesting to note that such a model can account for apparently irregular growth of the total sum $N(t)$.

5.4.2 Color Theory

Children who play with paintboxes soon discover that mixing yellow with blue produces green. Hence these colors are not *independent*. Thus instead of buying lots of paint tubes, one may consider that one tube of yellow and one tube of blue is equivalent to two tubes of green. (Note that this is not a strict equivalence: The entropy increases during the mixing, and it is not easy to separate green into its two components!) More precisely, one may form different combinations of blue B and yellow Y, written as *linear combinations*

$$B + Y = 2G: \text{ green}, \quad B + 2Y = 3G': \text{ spring green, etc.}$$

In short, we are building a model of color mixing within a vector space, using positive linear combinations of paints. In this space,

$$B + J \text{ and } 2B + 2J$$

lead to $2V$ and resp. $4V$, corresponding to the same *hue* (produced in double quantity in the second case). If hue is the only quality that interests us, we can limit ourselves in positive linear combinations having coefficients summing to 1, hence producing a unit quantity of desired color.

Available commercial paintboxes contain a set of *generators* for all colors, or so we hope! But as already noted, the dozen of paint tubes that they may contain are by no means independent. The relations

$$B + Y = 2G, \quad B + 2Y = 3G'$$

can even be written as *linear dependence relations*

$$B + Y - 2G = 0, \quad B + 2Y - 3G' = 0.$$

Physical mixtures correspond to linear combinations having nonnegative coefficients, but linear dependence relations have negative coefficients (in fact, the sum of the coefficients in a linear dependence relation is zero).

If we tend to be good with money, we may be looking for a minimal set of generators. Experience shows that blue and yellow are *independent colors*. We can complete these into a *basis* for all colors by adding red R.

Blue Red Yellow

Mixing the three in various proportions produces browns, or even black. In this naive theory, we consider that all hues may be obtained in the form

$$aR + bB + cY, \quad a \geqslant 0,\ b \geqslant 0,\ c \geqslant 0, \quad a + b + c = 1,$$

and conclude that our model vector space is three-dimensional, with *basis* R, Y, and B. This is the RYB color model for visual paint mixing. Pure hues are placed inside a triangle having vertices at the points R, Y and B. This triangle is reproduced in all "Teach yourself with Painting" books.

To describe the *grey value* of colors, we may add one dimension—represented by a diluter D—and thus consider linear combinations

$$aR + bB + cY + dD \text{ where } a + b + c + d = 1,$$

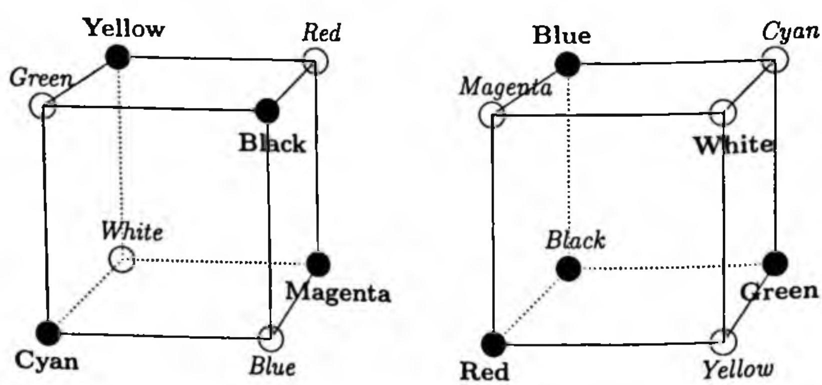
corresponding to the hue

$$aR + bB + cJ \quad (a + b + c \leqslant 1)$$

(all coefficients remaining nonnegative). To recover the first description, we have to cancel D. The original description is the projection onto the three-dimensional subspace generated by B, R, and Y, parallel to the direction generated by the diluter D. The enriched description (with grey value) can be envisioned as a pyramid having for basis the preceding triangle of colors, and vertex D (projecting onto 0).

The preceding visual theory for paint mixing is a *subtractive color theory* where the mixing of two paints absorbs the corresponding wavelengths, hence the resulting visible color is the remaining reflected spectrum. A physical theory for light-absorbing systems is based on the *primary colors Cyan* (greenish blue), *Magenta* (pinkish red), and *Yellow*. It is used in the process of color printing. Quadrichromic printers use the $CMYB$ colors, with *Black* added (cheaper than mixing complementary colors).

In an *additive theory*, superposition of colored light rays produces completely different results (e.g. mixing all colors produces white). The RGB additive theory has basis *Red*, *Green*, and *Blue*. It is well adapted for the interpretation of the behavior of colored computer screens. Handling images with computers, and printing them, emphasizes the differences of the two theories. Here are pictures exhibiting the differences (and relations) between them.



CMY subtractive model **RGB additive model**

There is no absolute model for the understanding of a natural phenomenon. Each one is suited to a special purpose, and constitutes an idealization of one aspect of this phenomenon. To illustrate this further, let us show how linear algebra can be used in another context linked with light propagation.

Polarization

When using a polarizing filter in photography, one has to double the exposure time. Hence such a filter absorbs 50% of the incoming light. However, two superposed polarizing filters do not attenuate the luminosity by a factor $\frac{1}{4} = \left(\frac{1}{2}\right)^2$, in general. The relative orientation of the filters is crucial, showing that the scalar interpretation of the attenuation is not suitable in this context. The following experiment shows that the introduction of a third polarizing filter can even increase the transmission factor!

Fix two perpendicular polarizing filters in a light ray so that no light can go through. Introduce a third polarizing filter between them, whose polarizing direction makes an angle of 45^0 with each: Observe now that some light goes through.

The Maxwell theory of light with electromagnetic waves gives a satisfactory explanation. In this theory, light is constituted by a pair of orthogonal vectors $\vec{E}, \vec{H}$, oscillating in a direction perpendicular to propagation. A polarizing filter only selects one component of these fields in a specific direction. Let us fix our attention to the electric field: A first polarizing filter with vertical orientation selects the vertical component of $\vec{E}$. A second orthogonal polarizing filter will select the horizontal component of $\vec{E}$. When both act in a sequence, no light can go through. But when the third intermediate filter is introduced, selecting the diagonal component, it is easy to understand that some light is transmitted. The amplitude of the initially vertically polarized ray is halved after going through the next two polarizing filters. Its intensity (the square of the amplitude) experiences an attenuation factor of $\frac{1}{4}$. This vector description of light waves makes it natural to use matrices in this context. The intermediate filter corresponds to a diagonal projection of the electric field $\vec{E}$ having matrix

$$P = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}.$$

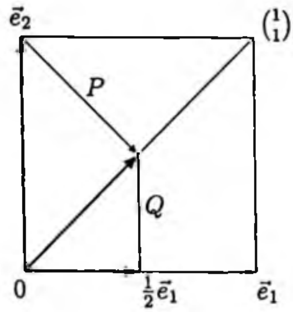
The last horizontally polarizing filter is a vertical projection with matrix

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

We have $Q(\vec{e}_2) = 0$ so that no light goes through with two orthogonal filters, while

$$Q \circ P = \begin{pmatrix} 0 & 0 \\ 1/2 & 1/2 \end{pmatrix}, \quad Q\,P(\vec{e}_2) = \tfrac{1}{2}\vec{e}_1.$$

Here is a picture of the attenuation produced by three successive polarizing filters with respective angles of 45^0.



5.4.3 Genetics

It is known that genes—located in chromosomes—are responsible for the transmission of characters in the reproduction process. *Alleles* are pairs of genes located at the same level, or *locus*, of a pair of chromosomes. A genotype is a pair of alleles at a locus of a chromosomic pair. Let us examine one locus, with two possible alleles A and a occurring with the respective frequencies p and q $(p + q = 1)$ in a given population.

1) If a father has the genotype AA, a descendent will inherit a gene A from his father and a gene A (probability p) or a (probability q) from his mother, so that his genotype will be

$$AA \text{ with probability } p, \qquad Aa \text{ with probability } q.$$

2) Let us determine the genotypes of the the descendents of a father having genotype Aa. There is equiprobability that the son inherits A or a from the father. The second character, inherited from the mother, is independent from the first one. Here is a table summing up the possibilities, with their probabilities.

$$Aa \begin{cases} A\,(\frac{1}{2}) & \rightarrow & AA\,(\frac{1}{2}p), & Aa\,(\frac{1}{2}q), & aa\,(0) \\ a\,(\frac{1}{2}) & \rightarrow & AA\,(0), & Aa\,(\frac{1}{2}p), & aa\,(\frac{1}{2}q) \\ \textit{Total:} & & AA\,(\frac{1}{2}p), & Aa\,(\frac{1}{2}), & aa\,(\frac{1}{2}q) \end{cases}$$

3) Finally, if the genotype of the father is aa, the genotype of the son will be

$$Aa \text{ with probability } p, \qquad aa \text{ with probability } q.$$

This explains the following complete table of possibilities

genotype of father ↓	genotype of son AA	Aa	aa
AA	p	q	0
Aa	$\frac{1}{2}p$	$\frac{1}{2}$	$\frac{1}{2}q$
aa	0	p	q

which we view as a *transition matrix*

$$T = \begin{pmatrix} p & q & 0 \\ \frac{1}{2}p & \frac{1}{2} & \frac{1}{2}q \\ 0 & p & q \end{pmatrix}.$$

The possibilities for a two-generations jump are determined by iteration. If a grandfather has genotype AA, his son will have genotypes AA with probability p, Aa with probability q, and aa with probability 0. The possibilities for the grandson are gathered in the following table.

$$AA \left\{ \begin{array}{llllll} AA\,(p) & \rightarrow & AA\,(p^2) & Aa\,(pq) & aa\,(p) \\ Aa\,(q) & \rightarrow & AA\,(\frac{1}{2}qp) & Aa\,(\frac{1}{2}q) & aa\,(\frac{1}{2}q^2) \\ aa\,(0) & \rightarrow & AA\,(0) & Aa\,(0) & aa\,(0). \end{array} \right.$$

Paths starting at AA and arriving at AA have a probability $p^2 + \frac{1}{2}pq$. We recognize a typical coefficient of the matrix product T^2.

5.4.4 Einstein Summation Convention

Let us take a basis $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ of $\mathbf{R}^n$. In order to avoid double indices, let us write the components of a basis vector in upper position

$$\mathbf{e}_j = \begin{pmatrix} e_j^1 \\ e_j^2 \\ \vdots \\ e_j^n \end{pmatrix}.$$

Since we work in linear algebra, there is little risk of confusion with powers. If necessary, one may denote the square of the component e_j^i as $(e_j^i)^2$ to avoid confusion between upper indices and powers. To be consistent, we shall also denote by upper indices all components of vectors. For example, we shall write $\mathbf{x} = \sum_{1 \leqslant i \leqslant n} x^i \mathbf{e}_i$.

CONVENTION. *When a summation concerns an index which appears both in upper and lower position, we delete it.*

Consequently, the above two formulas will here simply be written

$$\mathbf{x} = x^i \mathbf{e}_i = (\mathbf{e}_1, \dots, \mathbf{e}_n) \begin{pmatrix} x^1 \\ x^2 \\ \vdots \\ x^n \end{pmatrix}$$

(block multiplication). In order to be able to use this convention as often as possible (and delete all—or nearly all—summation signs), we have to be careful on the placement of indices. As we have already seen, the row index in a matrix should be placed in upper position (it refers to the components of the column vectors). Hence we denote by $A = (a^i_j)$ a typical matrix. Let us consider the product $C = AB$ of two matrices (of compatible sizes). If $A = (a^i_j)$ and $B = (b^k_\ell)$, a typical entry of C is

$$c^i_\ell = a^i_j b^j_\ell = a^i_k b^k_\ell$$

where the appearance of the index j in both positions suggests a summation on this index: The result is independent of the name of this index. Formally, we can say that an index placed in both positions cancels out in the result, just as in a fraction. The unit matrix I_n is the square matrix (δ^i_j) of size $n \times n$ and entries

$$\delta^i_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases}$$

(The indices of the Kronecker symbol now share an adequate position.) This unit matrix is characterized by the formulas

$$\delta^i_j a^j_k = a^i_k \quad \text{and} \quad a^i_j \delta^j_\ell = a^i_\ell.$$

In components, the formula $A\mathbf{x} = \mathbf{y}$ is now written $a^i_j x^j = y^i$ where the summation on the index j is implicitly made. As another example of this convention, the trace of a square matrix $A = (a^i_j)$ is simply written $\operatorname{tr} A = a^i_i$.

5.5 Exercises

1. Let

$$M = \begin{pmatrix} 1 & 1 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix}$$

What is the dimension of the kernel of M? Show that a multiple of M is a projector.

2. Using row operations, check that the rank of

$$A = \begin{pmatrix} 1 & 2 & 3 & \dots & n \\ 2 & 3 & 4 & \dots & n+1 \\ 3 & 4 & 5 & \dots & n+2 \end{pmatrix}$$

and of its transpose are the same.

3. Find a basis of the image of the linear map $\mathbf{R}^4 \to \mathbf{R}^4$ defined by the matrix

$$\begin{pmatrix} 1 & 1 & -1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & -1 & -1 & -1 \\ -1 & -1 & 1 & 0 \end{pmatrix}.$$

4. Let P be a projector in a vector space E, and define $\tau = \tau_P = I_E - 2P$. Check $\tau^2 = I_E$. With $Q = I_E - P$ (complementary projector), let also $\sigma = 2Q - I_E$, so that $\sigma^2 = I_E$. Show that in the usual 2-dimensional space, τ and σ are symmetries. What is $\tau \circ \sigma$? Any linear map $T : E \to E$ such that $T^2 = I_E$ is called a (generalized) *symmetry*, or an *involution*.

5. Let V_1 and V_2 be two vector subspaces of a vector space E. Prove that $V_1 \cap V_2$ is a vector subspace of E as follows. If $f_i : E \to F_i$ are linear maps with resp. kernels V_i $(i = 1, 2)$, consider the kernel of the linear map $g = (f_1, f_2)$ defined by

$$\begin{array}{ccc} E & \longrightarrow & F_1 \oplus F_2 \\ \mathbf{x} & \longmapsto & (f_1(\mathbf{x}), f_2(\mathbf{x})). \end{array}$$

6. Let E be the vector space of polynomials in two variables x and y, and

$$V = \mathcal{L}(x + y, (x \pm y)^2), \quad W = \mathcal{L}(1, x, y, xy).$$

What are the dimensions of V, W, $V + W$, and $V \cap W$?

7. (*a*) Let A and B be two matrices having the same number m of rows. How does one have to choose $\mathbf{b} \in \mathbf{R}^m$ in order that both $A\mathbf{x} = \mathbf{b}$ and $B\mathbf{y} = \mathbf{b}$ are compatible?

(*b*) Find the intersection of the images of the following matrices

$$A = \begin{pmatrix} 0 & 1 & 2 \\ 1 & 1 & 1 \\ 2 & 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \\ -1 & 1 & 0 \end{pmatrix}$$

8. (A Leslie Matrix) Consider a subdivision of a population into three age groups, each containing a generation of roughly 30 years:

$$x_1\text{: \# of young,} \quad x_2\text{: \# of workers,} \quad x_3\text{: \# of retired.}$$

Assume that the fertility of the second age group is 2/3 (about 1.3 children per couple) and that the survival ratio from the first group to the second group is 66% (due to illnesses, accidents, etc.) and 80% for the transition from second to third age group. We obtain the transition equations

$$x_1' = \tfrac{2}{3}x_2, \quad x_2' = \tfrac{2}{3}x_1, \quad x_3' = \tfrac{4}{5}x_2.$$

Show that the transition matrix admits ${}^t(5, 5, 6)$ as an eigenvector: Which is the corresponding eigenvalue? The corresponding pyramid of ages is stable: Show that it decreases steadily like $(2/3)^n = e^{-\alpha n}$ $\quad(\alpha = \log 3/2)$.

9. Consider the Leslie matrix A of size 9×9 corresponding to fertilities $f_i = 0$ $(i \neq 2, 3)$, $f_2 = 1/2$, $f_3 = 1/4$, and transition probabilities $p_0 = 69\%$, $p_1 = 61\%$, $p_2 = 51\%$, $p_3 = 42\%$, $p_4 = 33\%$,$p_5 = 24\%$, $p_6 = 16\%$, $p_7 = 10\%$, $p_8 = 0$. Write down the transition matrix A. Compute the sequence of pyramid of ages Av, A^2v, $\ldots$, $A^{10}v$ corresponding to the initial condition

$$v = {}^t(10^6, 0, 0, 0, 0, 0, 0, 0, 0).$$

Show in particular that the total population decreases twice, increases once, and then decreases again.

10. Let V be a vector subspace of a space E. The set of *affine subspaces* $x + V$ $(x \in E)$ parallel to V is a vector space if we define

$$\begin{aligned} a(x+V) &= ax + V, \\ (x+V) + (y+V) &= (x+y) + V, \end{aligned}$$

when a is a scalar, and x, $y \in E$. This vector space is called *quotient of E by V* and denoted by E/V. Show that the map

$$x \longmapsto x + V \ : \ E \longrightarrow E/V,$$

is linear, surjective, with kernel V. Show that for any surjective linear $f : E \to F$ with $\ker f = V$, F is isomorphic to E/V. When E is finite dimensional, show $\dim E/V = \dim E - \dim V$. Let $E = \mathbf{R}^n \times \mathbf{R}^m$ and $V = \{0\} \times \mathbf{R}^m \subset E$. What is E/V?

11. A surjective linear map $\pi : E \to F$ is called a *quotient map*. Show that for any quotient map π there is a linear map $\sigma : F \to E$ such that $\pi \circ \sigma = \mathrm{id}_F$ (use a basis of F to define σ). If $V = \ker \pi$ and $W = \mathrm{im}\,\sigma$, show that E is the direct sum of V and W (for $x \in E$, notice that $x - \sigma\pi x \in \ker \pi$). For each choice of σ, the composite $F \xrightarrow{\sigma} E \longrightarrow E/V$ (see previous exercise) is an isomorphism.

12. Let $W \subset V \subset E$ be vector subspaces of a space E. The canonical map

$$x + W \longmapsto x + V \quad : \quad E/W \longrightarrow E/V$$

is linear. Show that its kernel is $V/W \subset E/W$. Conclude that the quotient of E/W by V/W is isomorphic to E/V.

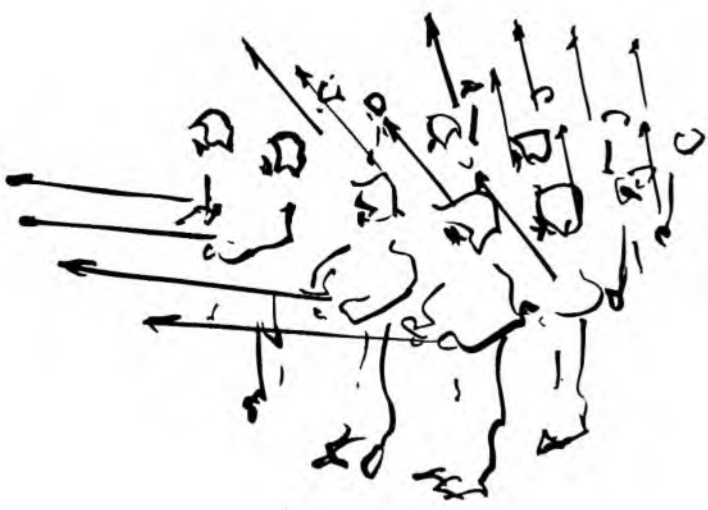
Notes

The method (Sec. 5.2.3) (construction of a basis of the intersection of two subspaces) comes from the article by Kung-Wei Yang

The Mathematical Magazine **70**, nb.4 (1997), p. 297.

Keywords for Web Search

Color theory, RGB, CMY
www.bway.net/ jscruggs/add.html
www.photoshopfocus.com/tips.htm
Mendelian genetics, Punnett square



Vectors were also used in the past!

Chapter 6

Eigenvectors and Eigenvalues

Here is the most important theoretical application of the rank theory. A whole chapter is devoted to the explanation of this concept, and a first section concentrates on one of its particularly simple aspects.

6.1 Introduction

Some people smoke, some do not. All know that it is a hazard for your health to do so... and some try to stop, or convince their children not to start, with various degrees of success. How will the proportion of smokers/nonsmokers vary, in a simple model where

among smokers, $\frac{3}{5}$ of their children smoke,

while

only $\frac{1}{5}$ of nonsmokers descendants start smoking?

Let us introduce a subdivision of the population into generations (taking for granted that the preceding proportions remain constant in future generations). Thus we denote by S and N the distribution at a given time and by S', N' the distribution one generation later. Our assumption can be translated into the equations

$$S' = \tfrac{3}{5}S + \tfrac{1}{5}N, \quad N' = \tfrac{2}{5}S + \tfrac{4}{5}N.$$

The matrix formalism allows us to rewrite these equations in the form

$$\begin{pmatrix} S' \\ N' \end{pmatrix} = \begin{pmatrix} \frac{3}{5} & \frac{1}{5} \\ \frac{2}{5} & \frac{4}{5} \end{pmatrix} \begin{pmatrix} S \\ N \end{pmatrix} = A \begin{pmatrix} S \\ N \end{pmatrix}.$$

The matrices A, A^2, A^3... furnish the dynamics of the evolution after one, two, three... generations. An interesting problem is to find one stable proportion (perhaps a limit one): For example, a train company might be interested in the proportion of smoking and nonsmoking carriages that it has to dispose in its network. In our particular example, let us check that the proportion $\frac{1}{3}$, $\frac{2}{3}$ is stable

$$\begin{pmatrix} \frac{3}{5} & \frac{1}{5} \\ \frac{2}{5} & \frac{4}{5} \end{pmatrix} \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \end{pmatrix} = \begin{pmatrix} \frac{3}{15} + \frac{2}{15} \\ \frac{2}{15} + \frac{8}{15} \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \end{pmatrix}.$$

In the same vein, the breakdown of a population according to age group can be represented by a vector $\mathbf{v}$. It may be possible to describe the evolution of this vector by a matrix multiplication: $\mathbf{v}' = A\mathbf{v}$. If the population changes, one may be especially concerned by the variation (or preservation) of the shape of the pyramid of ages. In the transition from one generation to the next one, this shape is preserved precisely when the vector representing the population is simply multiplied by a scalar $\mathbf{v}' = \lambda\mathbf{v}$. In other words, assuming that we know the matrix A, can we find a stable pyramid of ages? This problem leads to the theory which is explained below.

6.2 Definitions and Examples

Let us call *operator* any linear map from a vector space E *into itself.*

6.2.1 Definitions

Definition. *An* eigenvector *of an operator T is an element $\mathbf{v} \in E$ such that*

$$\mathbf{v} \neq 0 \quad \textit{and} \quad T\mathbf{v} \textit{ is proportional to } \mathbf{v}.$$

We then write $T\mathbf{v} = \lambda\mathbf{v}$ with a scalar λ. A nonzero vector $\mathbf{v}$ is an eigenvector of T when

$$T\mathbf{v} = \lambda\mathbf{v}, \quad T\mathbf{v} - \lambda\mathbf{v} = 0, \quad (T - \lambda I)\mathbf{v} = 0, \quad \mathbf{v} \in \ker(T - \lambda I).$$

Definition. *An* eigenvalue *of an operator T is a scalar λ such that*

$$\ker(T - \lambda I) \neq \{0\}.$$

The nonzero elements of $\ker(T - \lambda I)$ are the eigenvectors of T corresponding to the eigenvalue λ. The eigenvalues are the special values of a variable x such that $\ker(T - xI) \neq \{0\}$. When E is a finite-dimensional space, these are the values of x such that the rank of $T - xI$ is not maximal.

Definition. *If λ is an eigenvalue of an operator T in a vector space E,*

$$V_\lambda = \{\mathbf{v} \in E : T\mathbf{v} = \lambda\mathbf{v}\} = \ker(T - \lambda I)$$

is the eigenspace *of T relative to the eigenvalue λ. Its dimension*

$$m_\lambda = \dim V_\lambda = \dim \ker(T - \lambda I)$$

is the geometric multiplicity *of the eigenvalue* λ.

By definition

$$\lambda \textit{ eigenvalue of } T \iff V_\lambda = \ker(T - \lambda I) \neq \{0\} \iff m_\lambda \geqslant 1,$$

and the geometric multiplicity of λ is the maximal number of *linearly independent* eigenvectors that can be found for this eigenvalue. The rank theory explained in the first chapter allows us to compute this dimension by means of row operations: If $r = \operatorname{rank}(T - \lambda I)$, then this dimension is the number of free variables

$$m_\lambda = \dim \ker(T - \lambda I) = n - r.$$

6.2.2 Simple 2 × 2 Examples

A Special Method for Dimension 2

In the introduction to this chapter, we have encountered the matrix

$$\begin{pmatrix} \frac{3}{5} & \frac{1}{5} \\ \frac{2}{5} & \frac{4}{5} \end{pmatrix}.$$

Let us determine its eigenvectors. Since $\mathbf{e}_1 = \binom{1}{0}$ is not an eigenvector (its image is the first column of the matrix), any eigenvector will have a nonzero second component. Hence we may only look for eigenvectors having a normalized second component: We are looking for eigenvectors of the special form $\binom{x}{1}$, with second component equal to 1. The problem now is simply to find a first component x with

$$\begin{pmatrix} \frac{3}{5} & \frac{1}{5} \\ \frac{2}{5} & \frac{4}{5} \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix} = \lambda \begin{pmatrix} x \\ 1 \end{pmatrix}.$$

This condition is

$$\begin{pmatrix} \frac{3}{5}x + \frac{1}{5} \\ \frac{2}{5}x + \frac{4}{5} \end{pmatrix} = \begin{pmatrix} \lambda x \\ \lambda \end{pmatrix},$$

and leads to two equations for the pair λ and x

$$\begin{cases} \frac{3}{5}x + \frac{1}{5} & = & \lambda x \\ \frac{2}{5}x + \frac{4}{5} & = & \lambda. \end{cases}$$

Eliminating λ we find the condition

$$\tfrac{3}{5}x + \tfrac{1}{5} = \left(\tfrac{2}{5}x + \tfrac{4}{5}\right) x,$$

or

$$\tfrac{2}{5}x^2 + \tfrac{4}{5}x - \tfrac{3}{5}x - \tfrac{1}{5} = 0, \quad 2x^2 + x - 1 = 0.$$

The roots of this quadratic equation are

$$x = \frac{-1 \pm \sqrt{1+8}}{4} = \begin{cases} \frac{1}{2} \\ -1. \end{cases}$$

(1) A first eigenvector is obtained by taking $x = \frac{1}{2}$: $\vec{v} = \binom{1/2}{1}$. The corresponding eigenvalue is

$$\lambda = \tfrac{2}{5}x + \tfrac{4}{5} = \tfrac{2}{5}\tfrac{1}{2} + \tfrac{4}{5} = 1.$$

All nonzero multiples of this eigenvector $\binom{1/2}{1}$ are also eigenvectors with respect to the same eigenvalue. In particular, if we prefer integral components, we might take $\binom{1}{2}$, and if we prefer a sum of components equal to 1 as in the introduction, we would choose $\binom{1/3}{2/3}$.

(2) A second eigenvector is obtained by taking $x = -1$: $\vec{w} = \binom{-1}{1}$. The corresponding eigenvalue is

$$\mu = \tfrac{2}{5}(-1) + \tfrac{4}{5} = \tfrac{2}{5}.$$

This method is restricted to the 2-dimensional case. But in this case, it is elementary and very effective.

A Method for Dimension 2

Let us determine the eigenvectors of the matrix

$$A = \begin{pmatrix} 4 & -5 \\ 2 & -3 \end{pmatrix}.$$

Recall that they are the nonzero vectors $\vec{v}$ such that

$$A\vec{v} = \lambda\vec{v}, \quad A\vec{v} - \lambda\vec{v} = 0, \quad (A - \lambda I)\vec{v} = 0.$$

We can find such vectors provided $\ker(A - \lambda I) \neq \{0\}$, namely when $A - \lambda I$ is not injective, or equivalently not regular. Thus we determine first the values of a variable x such that the rank of $A - xI$ is not maximal. Let us proceed systematically with row operations

$$\begin{pmatrix} 4-x & -5 \\ 2 & -3-x \end{pmatrix} \sim \begin{pmatrix} 2 & -3-x \\ 4-x & -5 \end{pmatrix}$$

$$\sim \begin{pmatrix} 2 & -3-x \\ 0 & -5 - \frac{1}{2}(x-4)(x+3) \end{pmatrix}.$$

The rank is not maximal when

$$\begin{aligned} 2\left(5 + \tfrac{1}{2}(x-4)(x+3)\right) &= 10 + (x-4)(x+3) \\ &= x^2 - x - 2 \\ &= (x-2)(x+1) \end{aligned}$$

vanishes. The corresponding eigenvectors will have components satisfying

$$2v_1 - (3+\lambda)v_2 = 0.$$

We can take

$$\begin{aligned} v_2 &= 2 \\ v_1 &= 3+\lambda, \end{aligned}$$

and hence

$$\begin{aligned} \vec{v} &= \begin{pmatrix} 2 \\ 2 \end{pmatrix} \quad \text{eigenvector relative to } \lambda = -1, \\ \vec{w} &= \begin{pmatrix} 5 \\ 2 \end{pmatrix} \quad \text{eigenvector relative to } \lambda = 2. \end{aligned}$$

Observe that row operations furnish simultaneously the eigenvalues and the corresponding eigenvectors. (The reader should treat this matrix with the first particular method and compare them.)

6.2.3 A 4×4 Example

As a less trivial example, let us find the eigenvectors of the following 4×4 matrix

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

We proceed with row operations in order to find simultaneously the eigenvalues and the eigenvectors:

$$\begin{aligned} A - xI &= \begin{pmatrix} -x & 2 & -2 & 0 \\ 1 & 1-x & 0 & -1 \\ -1 & 1 & -2-x & 1 \\ -1 & 1 & -2 & 1-x \end{pmatrix} \\ &\sim \begin{pmatrix} 1 & 1-x & 0 & -1 \\ -x & 2 & -2 & 0 \\ -1 & 1 & -2-x & 1 \\ -1 & 1 & -2 & 1-x \end{pmatrix} \begin{array}{l} \\ \ell_2 + x\ell_1 \\ \ell_3 + \ell_1 \\ \ell_4 + \ell_1 \end{array} \\ &\sim \begin{pmatrix} 1 & 1-x & 0 & -1 \\ 0 & 2+x-x^2 & -2 & -x \\ 0 & 2-x & -2-x & 0 \\ 0 & 2-x & -2 & -x \end{pmatrix} \end{aligned}$$

$$A - xI \sim \begin{pmatrix} 1 & 1-x & 0 & -1 \\ 0 & 2-x & -2 & -x \\ 0 & 2-x & -2-x & 0 \\ 0 & 2+x-x^2 & -2 & -x \end{pmatrix} \begin{matrix} \\ \\ \ell_3 - \ell_2 \\ \ell_4 - (x+1)\ell_2 \end{matrix}$$

$$\sim \begin{pmatrix} 1 & 1-x & 0 & -1 \\ 0 & 2-x & -2 & -x \\ 0 & 0 & -x & +x \\ 0 & 0 & -2+2x+2 & -x+x^2+x \end{pmatrix}$$

$$\sim \begin{pmatrix} 1 & 1-x & 0 & -1 \\ 0 & 2-x & -2 & -x \\ 0 & 0 & -x & +x \\ 0 & 0 & 2x & x^2 \end{pmatrix} \begin{matrix} \\ \\ \\ \ell_4 + 2\ell_3 \end{matrix} .$$

Finally

$$A - xI \sim \begin{pmatrix} 1 & 1-x & 0 & -1 \\ 0 & 2-x & -2 & -x \\ 0 & 0 & -x & +x \\ 0 & 0 & 0 & x^2+2x \end{pmatrix} .$$

The rank is 4 except in the following cases

$x = \pm 2$: The rank is 3,
$x = 0$: The rank is 2.

Let now λ be one of these three eigenvalues. To find the corresponding eigenvectors, we solve the triangular homogeneous system

$$\begin{cases} x_1 + (1-\lambda)x_2 - x_4 = 0 \\ (2-\lambda)x_2 - 2x_3 - \lambda x_4 = 0 \\ - \lambda x_3 + \lambda x_4 = 0 \\ + (\lambda^2 + 2\lambda)x_4 = 0. \end{cases}$$

The following four eigenvectors make up a basis of the whole space

$$\begin{pmatrix} 1 \\ -1 \\ -1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \qquad \text{for } \lambda = 0 \text{ (in } \ker A),$$

$$\begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \text{ for } \lambda = 2, \qquad \begin{pmatrix} 1 \\ 0 \\ 1 \\ 1 \end{pmatrix} \text{ for } \lambda = -2.$$

6.2.4 Abstract Examples

(1) A *symmetry* is by definition an operator S satisfying $S^2 = \mathrm{id}$. Without any special property of the vector space E on which such an operator is defined, for any eigenvector $\mathbf{v}$ ($\neq 0$), we have

$$\begin{aligned} S\mathbf{v} = \lambda\mathbf{v} \quad &\Longrightarrow \quad S^2\mathbf{v} = S(\lambda\mathbf{v}) = \lambda S\mathbf{v} = \lambda^2\mathbf{v} \\ &\overset{S^2=\mathrm{id}}{\Longrightarrow} \quad \lambda^2 = 1: \quad \lambda = \pm 1. \end{aligned}$$

The only possible eigenvalues of a symmetry operator are ± 1. Here are special cases of this situation.

(1*a*) Let $E = M_n$ be the space of square matrices of size $n \times n$. Look at the transposition operator $S : A \mapsto {}^tA$. Obviously $S^2 = \mathrm{id}$ so that transposition is a symmetry. The eigenvectors of transposition are now the matrices A for which ${}^tA = \pm A$. For the eigenvalue $+1$, the eigenvectors are the matrices ${}^tA = A$, namely the *symmetric matrices*. For the eigenvalue -1, the eigenvectors are the matrices ${}^tA = -A$, namely the *skew-symmetric matrices*. Note that any (square) matrix A is a sum of a symmetric one A_s and a skew-symmetric one A_a. Indeed, for any square matrix A, $A + {}^tA$ is symmetric, while $A - {}^tA$ is skew-symmetric, and

$$A = \underbrace{\tfrac{1}{2}(A + {}^tA)}_{=A_s} + \underbrace{\tfrac{1}{2}(A - {}^tA)}_{=A_a}.$$

In fact, this is the only decomposition of A as such a sum: Suppose $A = X + Y$ where X is symmetric and Y is skew-symmetric. Then

$$\begin{aligned} A &= X + Y, \\ {}^tA &= {}^tX + {}^tY = X - Y, \end{aligned}$$

whence $A + {}^tA = 2X$, $A - {}^tA = 2Y$ and necessarily

$$X = \tfrac{1}{2}(A + {}^tA), \quad Y = \tfrac{1}{2}(A - {}^tA).$$

It is reasonable to call

$$\begin{aligned} X &= \tfrac{1}{2}(A + {}^tA) = A_s \quad \textit{the symmetric part of } A, \\ Y &= \tfrac{1}{2}(A - {}^tA) = A_a \quad \textit{the skew-symmetric part of } A. \end{aligned}$$

(1*b*) Let E be a vector space of functions on the real line, such that if a function f is in E, its symmetric part f_s defined by $f_s(x) = f(-x)$ is also in E. Define an operator $S : E \to E$ by $Sf = f_s$. Obviously $S^2 f = f$ so that this operator is a *symmetry*. The eigenfunctions corresponding to the eigenvalue $+1$ are the solutions of $Sf = f$, namely the functions f satisfying $f(-x) = f(x)$ identically. These functions are called *even functions*. The eigenfunctions corresponding to the eigenvalue -1 are the *odd functions*, characterized by $f(-x) = -f(x)$

(identically). As in the preceding example (1a), any function f can be written in a unique way as a sum of an even and an odd function

$$f(x) = \underbrace{\tfrac{1}{2}(f(x) + f(-x))}_{\text{even}} + \underbrace{\tfrac{1}{2}(f(x) - f(-x))}_{\text{odd}}.$$

For a polynomial $\sum_{0 \leqslant i \leqslant n} a_i x^i$, the even part is

$$\sum_{0 \leqslant i = 2j \text{ even } \leqslant n} a_{2j} x^{2j},$$

while the odd part is

$$\sum_{0 \leqslant i = 2j+1 \text{ odd } \leqslant n} a_{2j+1} x^{2j+1}.$$

Another example is supplied by the exponential function:

$$\begin{aligned} \cosh x &= \tfrac{1}{2}(e^x + e^{-x}) \quad \text{is the } \textit{even part} \text{ of the exponential} \\ \sinh x &= \tfrac{1}{2}(e^x - e^{-x}) \quad \text{is the } \textit{odd part} \text{ of the exponential.} \end{aligned}$$

Hence the preceding decomposition for the exponential is

$$f(x) = e^x = \cosh x + \sinh x.$$

(2) The eigenvectors relative to the eigenvalue $\lambda = 0$ are the nonzero elements in the kernel of T. For example, let N be a *nilpotent* operator, say $N^k = 0$ for some integer $k > 1$, but $N^{k-1} \neq 0$. Then the image of N^{k-1} is contained in the kernel of N, hence consists of eigenvectors of N relative to the eigenvalue 0. For all $\mathbf{v} \neq 0$, the last nonzero element in the sequence $\mathbf{v}$, $N\mathbf{v}$, $N^2\mathbf{v}$,... is an eigenvector of N with eigenvalue 0.

(3) The eigenvectors relative to the eigenvalue $\lambda = 1$ are the nonzero fixed elements. Consider for example the unit translation operator on continuous functions. Here, T acts in $C(\mathbf{R})$ by

$$(Tf)(x) = f(x+1) \quad (f \in C(\mathbf{R})).$$

The eigenvectors relatively to the eigenvalue 1 are the functions f satisfying

$$f(x+1) = f(x) \quad (x \in \mathbf{R}).$$

These are the *periodic functions* of period one. For example, the functions

$$\sin 2\pi kx, \ \cos 2\pi kx \quad (k \geqslant 0)$$

are eigenvectors of the unit translation.

(4) In the vector space C^∞ of smooth (namely indefinitely differentiable) functions f on the real line, consider the derivation operator

$$D : f \longmapsto f' \quad (f \in C^\infty).$$

Which functions f have a derivative proportional to f? In other words, for which functions f is the derivative $f' = Df$ a multiple of f, say $Df = \lambda f$? The answer is well-known: These are the exponentials $f(t) = ce^{\lambda t}$. These are also eigenfunctions of the second derivative operator

$$D^2 : f \longmapsto f'', \quad D^2(e^{\lambda t}) = \lambda^2 e^{\lambda t}.$$

If we are looking for eigenfunctions of D in the subspace E consisting of the functions $f \in C^\infty$ satisfying $f(1) = f(0)$, then we have to select among the exponentials those for which $e^\lambda = e^0 = 1$: The only real case is $\lambda = 0$, and the corresponding eigenvectors are the constants. But if we are looking for eigenfunctions of D^2 in the same subspace, we may proceed as follows. For λ an integral multiple of the complex number $2\pi i$, say $\lambda = 2i\pi k$, the eigenvalue of D^2 is the real number $\lambda^2 = -4\pi^2 k^2 \leqslant 0$. The two complex eigenvectors $e^{\pm 2i\pi k}$ are independent, correspond to the same eigenvalue λ^2, and may be combined into the real functions

$$\cos 2\pi kt = \tfrac{1}{2}(e^{2i\pi kt} + e^{-2i\pi kt}), \quad \sin 2\pi kt = \tfrac{1}{2i}(e^{2i\pi kt} - e^{-2i\pi kt}),$$

which are thus *real eigenvectors of D^2 having nonpositive eigenvalues.*

(5) For any eigenvector $\mathbf{v}$ of an operator P, say $\mathbf{v} \neq 0$ and $P\mathbf{v} = \lambda\mathbf{v}$, we have

$$P^2\mathbf{v} = P(\lambda\mathbf{v}) = \lambda P\mathbf{v} = \lambda^2\mathbf{v}.$$

If P is a projector, namely $P^2 = P$, we infer $\lambda^2\mathbf{v} = \lambda\mathbf{v}$, hence $\lambda^2 = \lambda$ (since $\mathbf{v} \neq 0$). This proves that the only possible eigenvalues of a projector are $\lambda = 0$ or 1. The eigenspace corresponding to $\lambda = 0$ is the kernel of P while the eigenspace corresponding to $\lambda = 1$ is $\ker(I - P) = \operatorname{im} P$, as we have seen.

(6) If $\mathbf{v}$ is an eigenvector of an operator T, say $\mathbf{v} \neq 0$ and $T\mathbf{v} = \lambda\mathbf{v}$, then $T^n\mathbf{v} = \lambda^n\mathbf{v}$ ($n \geqslant 1$) and proceeding with linear combinations, $f(T)\mathbf{v} = f(\lambda)\mathbf{v}$ for any polynomial f. Hence $\mathbf{v}$ is an eigenvector of $f(T)$ relatively to the eigenvalue $f(\lambda)$. If $f(T) = 0$, then the eigenvalues are to be found among the zeros of f, namely the solutions of $f(\lambda) = 0$. Several cases have appeared

- $T^2 - I = 0$, T is a symmetry, the only possible eigenvalues are ± 1
- $T^2 - T = 0$, T is a projector, the only possible eigenvalues are 0 and 1
- $T^n = 0$ (for some integer $n \geqslant 1$), T is a nilpotent operator, the only possible eigenvalue is 0.

(7) Here is a useful trick for the construction of particularly simple examples. Let $A = (a_{ij})$ be a square matrix, and consider the column vector having all its components equal to 1. Then

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} s_1 \\ \vdots \\ s_n \end{pmatrix},$$

where $s_i = \sum_j a_{ij}$ is the sum of the entries in the ith row. If all row sums are the same, say $s_i = \sum_j a_{ij} = s$ for all i, we see that $\mathbf{v} = {}^t(1, 1, \ldots)$ is an eigenvector of A with respect to the eigenvalue $\lambda = s$. For example

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

Here is another example in dimension n:

$$\begin{pmatrix} a & b & \ldots & b \\ b & a & \ldots & b \\ \vdots & & \ddots & \vdots \\ b & b & \ldots & a \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = (a + (n-1)b) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

(8) Finally, let us quote the important example of the *Schrödinger equation*. The Hamiltonian of a physical system is a partial differential operator $\mathcal{H}$ acting on a certain space of functions. The (suitably normalized) eigenvectors of $\mathcal{H}$ are the *stationary states* ψ of the system. If $\mathcal{H}\psi = E\psi$, the eigenvalue E is the energy in the state ψ. The partial differential equation $\mathcal{H}\psi = E\psi$ is the Schrödinger equation of the system. For the hydrogen atom, there is a sequence of stationary states ψ_n corresponding to energy levels E_n: $\mathcal{H}\psi_n = E_n\psi_n$ $(n \geqslant 1)$.

6.3 General Results

6.3.1 Estimation of the Number of Eigenvalues

We intend to prove that an $n \times n$ matrix *has at most n eigenvalues.* This fundamental fact will be obtained as a consequence of an understanding of the relative position of the eigenspaces.

Recall that a family of subspaces V_i of a vector space E is *independent*, if a finite sum $\sum \mathbf{v}_i$ of elements $\mathbf{v}_i \in V_i$ can vanish only when all $\mathbf{v}_i = 0$ (Sec. 5.2.4). Quite generally, without finiteness assumption on the dimension, we have the following result.

Theorem. The eigenspaces of an operator $T : E \to E$ form an independent family of subspaces of E.

PROOF. Consider a finite sum $\sum \mathbf{v}_i = 0$ where $T\mathbf{v}_i = \lambda_i \mathbf{v}_i$, with distinct eigenvalues λ_i. We have to prove $\mathbf{v}_i = 0$ for all i. This is done by induction on the number of terms in this sum. If there is only one term, there is nothing to prove. The reduction of a sum to a shorter one is done as follows. Observe that the two relations

$$\begin{aligned} \sum \lambda_i \mathbf{v}_i &= T\Big(\sum \mathbf{v}_i\Big) &= 0, \\ \sum \lambda_1 \mathbf{v}_i &= \lambda_1 \sum \mathbf{v}_i &= 0, \end{aligned}$$

have the same first term. By subtraction, we obtain a shorter one

$$\sum (\lambda_i - \lambda_1)\mathbf{v}_i = 0.$$

By induction hypothesis, each term in this shorter relation vanishes

$$(\lambda_i - \lambda_1)\mathbf{v}_i = 0 \quad \text{for all } i.$$

Since the eigenvalues are distinct, we deduce $\mathbf{v}_i = 0$ ($i \geqslant 2$), and finally also $\mathbf{v}_1 = 0$. ∎

The crucial property valid for a family of independent subspaces is the following:

> *If we choose an independent subset S_i in each V_i $(i \in I)$, then the union $\bigcup_{i \in I} S_i \subset E$ is linearly independent.*

It immediately furnishes the important result quoted at the head of this section.

Corollary. *The sum of the geometric multiplicities of the eigenvalues is less than or equal to* $\dim E$. *In particular, if* $\dim E = n < \infty$, *then T has at most n distinct eigenvalues.* ∎

6.3.2 Localization of Eigenvalues

It is easy to localize the eigenvalues of a matrix with the Gershgorin condition for invertibility (Sec. 4.3.3).

Theorem (Gershgorin). *Let $A = (a_{ij}) \in M_n(\mathbf{R})$ be a square matrix, and define $r_i = \sum_{j \neq i} |a_{ij}|$ $(1 \leqslant i \leqslant n)$. Then the eigenvalues of A are contained in the union of the intervals $[a_{ii} - r_i, a_{ii} + r_i]$.*

PROOF. Let $\mathbf{v} \neq 0$ be an eigenvector of A, say $A\mathbf{v} = \lambda \mathbf{v}$. In components $\sum_j a_{ij} v_j = \lambda v_i$, and

$$(\lambda - a_{ii}) v_i = \sum_{j \neq i} a_{ij} v_j \quad (1 \leqslant i \leqslant n).$$

Choose an index i such that $|v_i|$ is maximal ($\neq 0$ since $\mathbf{v} \neq 0$), and divide by v_i

$$\begin{aligned} \lambda - a_{ii} &= \sum_{j \neq i} a_{ij} \frac{v_j}{v_i}, \\ |\lambda - a_{ii}| &\leqslant \sum_{j \neq i} |a_{ij}| \left| \frac{v_j}{v_i} \right| \\ &\leqslant \sum_{j \neq i} |a_{ij}| = r_i. \end{aligned}$$

This proves that λ is in the interval centered at a_{ii} and of length $2r_i$. ∎

Warning. Just as for the Gershgorin invertibility condition, the preceding statement is not valid for *any* field of scalars since it uses the absolute value. For the complex field $\mathbf{C}$, real intervals are to be replaced by complex discs (Sec. 11.2.2).

6.3.3 A Method for Finding Eigenvectors

Let A be a square matrix of size $n \times n$. The eigenvectors of A are the nonzero n-tuples $\mathbf{v} \in \mathbf{R}^n$ such that $A\mathbf{v}$ is proportional to $\mathbf{v}$. Here is a method for finding them. Start with *any* nonzero $\mathbf{v} \in \mathbf{R}^n$, and consider the sequence

$$\mathbf{v}, A\mathbf{v}, \ldots, A^k\mathbf{v}, \ldots$$

Let $k \leqslant n$ be the smallest integer such that $\mathbf{v}, A\mathbf{v}, \ldots, A^k\mathbf{v}$ is not independent. Hence there is a linear dependence relation having the form

$$A^k\mathbf{v} + a_{k-1}A^{k-1}\mathbf{v} + \cdots + a_1 A\mathbf{v} + a_0\mathbf{v} = 0.$$

Consider the corresponding polynomial

$$p(x) = x^k + a_{k-1}x^{k-1} + \cdots + a_1 x + a_0.$$

Any root of this polynomial leads to a factorization

$$p(x) = (x - \lambda)(x^{k-1} + b_{k-2}x^{k-2} + \cdots + b_1 x + b_0),$$

and hence to

$$(A - \lambda I)(A^{k-1}\mathbf{v} + b_{k-2}A^{k-2}\mathbf{v} + \cdots + b_1 A\mathbf{v} + b_0\mathbf{v}) = 0.$$

Since we are assuming that k is minimal, the vectors $\mathbf{v}, A\mathbf{v}, \ldots, A^{k-1}\mathbf{v}$ are linearly independent so that

$$A^{k-1}\mathbf{v} + b_{k-2}A^{k-2}\mathbf{v} + \cdots + b_1 A\mathbf{v} + b_0\mathbf{v} \neq 0.$$

This proves that this vector is an eigenvector of A relative to the eigenvalue λ. In this way, each root of the polynomial p leads to an eigenvector. (If the root λ is a complex number, the factorization has complex coefficients and we find an eigenvector with complex components.) With some luck on the choice of $\mathbf{v} \neq 0$ (some experience, accounting for a special form of the matrix A), one may find a small value of k, and thus find eigenvectors relatively easily. In general however, $k = n$ and there is no easy way of finding a factorization of this polynomial. (Evariste Galois (1811-1832) proved that there is no *algebraic formula* for finding the roots of an nth degree equation, when $n \geqslant 5$.)

Example. Let $A = \begin{pmatrix} 2 & 3 & -1 \\ 0 & 1 & 0 \\ -2 & 1 & 1 \end{pmatrix}$. Taking $\mathbf{v} = \mathbf{e}_1$, we find successively

$$\mathbf{v} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad A\mathbf{v} = \begin{pmatrix} 2 \\ 0 \\ -2 \end{pmatrix}, \quad A^2\mathbf{v} = \begin{pmatrix} 6 \\ 0 \\ -6 \end{pmatrix}.$$

Hence a linear dependence relation $A^2\mathbf{v} = 3A\mathbf{v}$ or equivalently $(A^2 - 3A)\mathbf{v} = 0$. The factorization $x^2 - 3x = x(x-3) = (x-3)x$ leads to two eigenvectors:
(1) $A(A-3I)\mathbf{v} = 0$ implies that $(A-3I)\mathbf{v}$ is an eigenvector of A relatively to the eigenvalue 0. Here is this eigenvector

$$(A-3I)\mathbf{v} = \begin{pmatrix} 2 \\ 0 \\ -2 \end{pmatrix} - \begin{pmatrix} 3 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ -2 \end{pmatrix} \in \ker A.$$

(2) $(A-3I)A\mathbf{v} = 0$ implies that $A\mathbf{v}$ is an eigenvector of A relatively to the eigenvalue 3. Here it is

$$A\mathbf{v} = \begin{pmatrix} 2 \\ 0 \\ -2 \end{pmatrix}.$$

(3) Starting with $\mathbf{v} = \mathbf{e}_2$, which is linearly independent from the previously found eigenvectors, one finds easily

$$A^3\mathbf{e}_2 = 4A^2\mathbf{e}_2 - 3A\mathbf{e}_2, \quad (A^3 - 4A^2 + 3A)\mathbf{e}_2 = 0.$$

The factorization $A(A-3I)(A-I)\mathbf{e}_2 = 0$ leads to a third eigenvector $A(A-3I)\mathbf{e}_2$ with respect to the third eigenvalue 1.

6.3.4 Eigenvectors and Commutation

Two operators S and T *commute* when $ST = TS$. When this is the case $S(\ker T) \subset \ker T$:

$$\mathbf{v} \in \ker T \implies T(S\mathbf{v}) = S\underbrace{T\mathbf{v}}_{=0} = 0 \iff S\mathbf{v} \in \ker T.$$

When S and T commute, S also commutes with $T - \lambda I$ and

$$S(\underbrace{\ker(T-\lambda I)}_{V_\lambda(T)}) \subset \underbrace{\ker(T-\lambda I)}_{V_\lambda(T)}.$$

Definition. *A family of operators $(S_i)_{i\in I}$ in a vector space E is called* irreducible *when the only subspaces $V \subset E$ such that $S_i(V) \subset V$ for all $i \in I$ are $V = \{0\}$ and $V = E$.*

A subspace $V \subset E$ such that $S_i(V) \subset V$ for all $i \in I$ is often called an *invariant subspace* of the family $(S_i)_{i\in I}$.

Lemma **(*Schur's lemma*).** *Let $(S_i)_{i\in I}$ be an* irreducible *family of operators in a vector space E. Then any operator T* having an eigenvector, *that commutes with all S_i is a multiple of the identity.*

PROOF. Let $\mathbf{v}$ be an eigenvector of T and λ its eigenvalue. Then the eigenspace $V_\lambda = \ker(T - \lambda I)$ is nonzero since it contains $\mathbf{v} \neq 0$. It is invariant under all S_i:

$$\mathbf{x} \in V_\lambda \iff (T - \lambda I)\mathbf{x} = 0 \implies S_i(T - \lambda I)\mathbf{x} = 0,$$

hence by the commutation property

$$(T - \lambda I)S_i\mathbf{x} = 0,$$

so that $S_i\mathbf{x} \in V_\lambda$. Since the family $(S_i)_{i \in I}$ is irreducible by assumption, this eigenspace V_λ is the whole space E: $T = \lambda I$. ∎

6.4 Applications of Eigenvectors

6.4.1 The Fibonacci Numbers

The Fibonacci sequence $(f_n)_{n \geqslant 0}$ is defined by

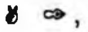
$$f_0 = 0, \ f_1 = 1 \text{ and } f_{n+1} = f_n + f_{n-1} \quad (n \geqslant 1).$$

Here is the beginning of this sequence

$$0, 1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, \ldots$$

We intend to show that it grows *exponentially*: For this purpose, we are going to give a formula for its general nth term.

This sequence arises for example if we consider the growth of rabbits, in a simple model, where a pair produces a new pair at each generation, admitting that fertility starts at the second generation. (Here, we disregard mortality, and admit that reproduction holds forever, a simplifying—but unreal—assumption.) For example, an initial pair does not reproduce in the first generation, being too young. It only starts at the second generation, producing a new pair. Let us represent old pairs by  and young ones by . At this stage, we have a family



and at the next stage

The old pair has reproduced, while the young one has matured. Still one generation later, we shall have

This is the way the Fibonacci sequence originally appeared. Each generation is constituted by a pyramid of ages having two components: The number of mature couples x_n, the number of young couples y_n. Let us make a vector with these components

$$\vec{f}_n = \begin{pmatrix} x_n \\ y_n \end{pmatrix} \mapsto f_n = x_n + y_n.$$

The number of pairs f_n is the sum of the number f_{n-1} of pairs which existed at the preceding generation and the number of young ones, produced by the old pairs, hence produced by the $x_n = f_{n-2}$ pairs which already existed two generations before

$$f_n = f_{n-1} + f_{n-2}.$$

These considerations suggest to use the matrix description for the evolution of the pyramid of ages of this population. Instead of f_n, we have a vector $\vec{f}_n = \binom{f_{n-1}}{f_{n-2}}$ with sum of components f_n. The transition is given by a 2×2 symmetric matrix

$$\vec{f}_{n+1} = A\vec{f}_n, \qquad \begin{pmatrix} f_n \\ f_{n-1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} f_{n-1} \\ f_{n-2} \end{pmatrix}.$$

By induction we see

$$\begin{pmatrix} f_{n+1} \\ f_n \end{pmatrix} = A^n \begin{pmatrix} f_1 \\ f_0 \end{pmatrix} = A^n \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

and hence

$$\text{1st column of } A^n = A^n \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} f_{n+1} \\ f_n \end{pmatrix}.$$

We see similarly that

$$\text{2nd column of } A^n = A^n \begin{pmatrix} 0 \\ 1 \end{pmatrix} = A^{n-1} \cdot A \begin{pmatrix} 0 \\ 1 \end{pmatrix} = A^{n-1} \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

so that this second column is the first column of A^{n-1}. It proves

$$A^n = \begin{pmatrix} f_{n+1} & f_n \\ f_n & f_{n-1} \end{pmatrix} \quad \text{(a symmetric matrix)}.$$

The explicit determination of A^n will give an explicit form for f_n.

Proposition. *Let $\lambda = \frac{1}{2}(1 + \sqrt{5})$, $\mu = \frac{1}{2}(1 - \sqrt{5}) = 1 - \lambda$ be the roots of the quadratic equation $x^2 = x + 1$. Then the nth Fibonacci number is*

$$f_n = \frac{1}{\sqrt{5}}(\lambda^n - \mu^n) = \textit{ integer closest to } \frac{1}{\sqrt{5}}\lambda^n.$$

PROOF. To compute the powers of the Fibonacci matrix A, we use a basis in which it is diagonal. For this, we determine the eigenvectors of A. Since the first basis vector is not an eigenvector, we may only consider vectors having a nonzero second component. Hence (normalization), we may only look for eigenvectors having a second component equal to 1. Here is the condition for these eigenvectors

$$A\begin{pmatrix} x \\ 1 \end{pmatrix} = \begin{pmatrix} x+1 \\ x \end{pmatrix} \quad \text{multiple of } \begin{pmatrix} x \\ 1 \end{pmatrix}.$$

Comparing the second components, the only possibility for the proportionality factor is x and comparing the first components, we see that $\binom{x}{1}$ is an eigenvector precisely when $x + 1 = x^2$. Since the quadratic equation $x^2 - x - 1 = 0$ has two real roots

$$\lambda = \tfrac{1}{2}(1+\sqrt{5}) = 1.618034\ldots, \qquad \mu = \tfrac{1}{2}(1-\sqrt{5}) = 1 - \lambda,$$

there are two independent eigenvectors. Let S be the matrix having for columns two such eigenvectors, say

$$S = \begin{pmatrix} \lambda & \mu \\ 1 & 1 \end{pmatrix}.$$

This matrix is invertible (the eigenvectors are independent) and

$$AS = A\left(\boxed{\begin{matrix} \lambda \\ 1 \end{matrix}}\ \boxed{\begin{matrix} \mu \\ 1 \end{matrix}}\right) = \left(\lambda\boxed{\begin{matrix} \lambda \\ 1 \end{matrix}}\ \mu\boxed{\begin{matrix} \mu \\ 1 \end{matrix}}\right) = \begin{pmatrix} \lambda & \mu \\ 1 & 1 \end{pmatrix}\begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix}$$

(a column operation is given by a product—at the right side—by an elementary matrix). We have found

$$S^{-1}AS = \begin{pmatrix} \lambda & 0 \\ 0 & \mu \end{pmatrix} = D: \text{ diagonal,}$$

or equivalently $A = SDS^{-1}$. Hence

$$A^n = SDS^{-1} \cdot SDS^{-1} \cdots SDS^{-1} = SD^nS^{-1}.$$

Since

$$S^{-1} = \frac{1}{\sqrt{5}}\begin{pmatrix} 1 & -\mu \\ -1 & \lambda \end{pmatrix},$$

we see that

$$\begin{aligned} \begin{pmatrix} f_{n+1} \\ f_n \end{pmatrix} &= 1st \text{ column of } A^n = A^n\begin{pmatrix} 1 \\ 0 \end{pmatrix} = SD^nS^{-1}\begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ &= \frac{1}{\sqrt{5}}\begin{pmatrix} \lambda & \mu \\ 1 & 1 \end{pmatrix} D^n \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \frac{1}{\sqrt{5}}\begin{pmatrix} \lambda & \mu \\ 1 & 1 \end{pmatrix}\begin{pmatrix} \lambda^n \\ -\mu^n \end{pmatrix}. \end{aligned}$$

We are interested in the second component

$$f_n = \frac{1}{\sqrt{5}}(\lambda^n - \mu^n).$$

This is the famous *Binet Formula.*[1] From $\left|\frac{1}{\sqrt{5}}\mu\right| < \frac{1}{2}$ we infer $\left|\frac{1}{\sqrt{5}}\mu^n\right| < \frac{1}{2}$

$$f_n = \text{ integer closest to } \frac{1}{\sqrt{5}}\lambda^n. \qquad \blacksquare$$

From it, we may deduce the speed of growth of the Fibonacci sequence: Its growth is *exponential*

$$f_n \approx \frac{1}{\sqrt{5}}\lambda^n = Ce^{\alpha n},$$

where $C = \frac{1}{\sqrt{5}} = 0,447\ldots$ and $\alpha = \log\lambda = 0,481\ldots$

Comment. The convergence $\frac{1}{\sqrt{5}}\mu^n \to 0$ is very fast, and the approximation $f_n \approx \frac{1}{\sqrt{5}}\lambda^n$ is excellent. Here are a few values

$$\begin{aligned}
\frac{1}{\sqrt{5}}\lambda^{10} &= 55.0036\ldots \approx f_{10} = 55,\\
\frac{1}{\sqrt{5}}\lambda^{20} &= 6765.00002956\ldots \approx f_{20} = 6765,\\
\frac{1}{\sqrt{5}}\lambda^{30} &= 832040.00000024037 \approx f_{30} = 832\,040,\\
f_{50} &= 12\,586\,269\,025.
\end{aligned}$$

6.4.2 Diagonalization

The operators which are given by a diagonal matrix in a basis have these basis elements as independent eigenvectors. Conversely, if a finite-dimensional space V has a basis consisting of eigenvectors of an operator T, then the matrix of T in this basis is diagonal with the eigenvalues as entries in the diagonal.

Consider a square matrix A of size $n \times n$ admitting n independent eigenvectors $\mathbf{v}_i \in \mathbf{R}^n$, say $A\mathbf{v}_i = \lambda_i\mathbf{v}_i$. The matrix S having the $\mathbf{v}_i$ for columns is regular and

$$\begin{aligned}
AS &= A\begin{pmatrix} \boxed{\mathbf{v}_1} & \cdots & \boxed{\mathbf{v}_n} \end{pmatrix} = \begin{pmatrix} \boxed{A\mathbf{v}_1} & \cdots & \boxed{A\mathbf{v}_n} \end{pmatrix}\\
&= \begin{pmatrix} \boxed{\lambda_1\mathbf{v}_1} & \cdots & \boxed{\lambda_n\mathbf{v}_n} \end{pmatrix} = S\begin{pmatrix} \lambda_1 & 0 & \cdots & 0\\ 0 & \ddots & & \vdots\\ \vdots & & \ddots & 0\\ 0 & \cdots & 0 & \lambda_n \end{pmatrix}.
\end{aligned}$$

[1] Jacques P.M. Binet (1786-1856).

This shows

$$A = S \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \ddots & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} S^{-1}.$$

Definition. *A square matrix A is* diagonalizable *if there is an invertible matrix S such that*

$$S^{-1}AS = D: \ \textit{diagonal matrix.}$$

Let $\mathbf{e}_j$ be the jth vector in the canonical basis of $\mathbf{R}^n$. Then $D\mathbf{e}_j = \lambda_j \mathbf{e}_j$ where λ_j is the entry of D placed in the jth row and jth column. From $S^{-1}AS = D$, we infer

$$\begin{aligned} S^{-1}AS\mathbf{e}_j &= D\mathbf{e}_j = \lambda_j \mathbf{e}_j, \\ AS\mathbf{e}_j &= SD\mathbf{e}_j = \lambda_j S\mathbf{e}_j, \end{aligned}$$

hence the jth column $S\mathbf{e}_j = \mathbf{v}_j$ of S is an eigenvector for this matrix with respect to the eigenvalue λ_j.

An operator T in a vector space V is *diagonalizable* if there exists a basis of V in which the matrix of T is diagonal.

Theorem. *Let T be an operator in a finite-dimensional real vector space V. Then T is diagonalizable precisely when V has a basis consisting of eigenvectors of T, hence precisely when*

$$\sum m_\lambda = \dim V,$$

a sum extended over the eigenvalues of T. ■

Corollary 1. *Any operator in an n-dimensional vector space E, having n distinct eigenvalues is diagonalizable.*

PROOF. Indeed, any system of eigenvectors corresponding to distinct eigenvalues is linearly independent. ■

Corollary 2. *Any triangular matrix having distinct entries in its diagonal is diagonalizable.*

PROOF. Each diagonal entry a_{ii} of a triangular matrix A is an eigenvalue: Indeed row operations show that the rank of $A - a_{ii}I$ is less than n, so that $\ker(A - a_{ii}I) \neq \{0\}$. Hence the assumption implies that A has n distinct eigenvalues. ■

Examples. (1) The 3×3 matrix

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{pmatrix},$$

has the three eigenvalues 1, 2, and 3, hence is diagonalizable.

(2) The matrix $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ cannot be put into diagonal form with real scalars. More generally, the matrices $\begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}$ cannot be put into diagonal form with real scalars (when φ is not an integral multiple of π). However, in the complex field they have two distinct eigenvalues $e^{\pm i\varphi}$, so that these matrices have diagonal forms over $\mathbf{C}$.

(3) The matrix $A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ has only one eigenvalue $\lambda = 1$. The corresponding eigenspace is $\ker(A-I) = \ker \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, a subspace of dimension 1. The geometric multiplicity of this eigenvalue is 1, so that A cannot be diagonalized.

(4) The matrices

$$A = \begin{pmatrix} a & b & \dots & b \\ b & a & \dots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \dots & a \end{pmatrix}$$

can be diagonalized. Indeed $A + (b-a)I_n = bM$ where

$$M = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix}$$

has rank 1, hence a kernel of dimension $n-1$ by the rank-nullity theorem. In any basis of $\mathbf{R}^n$ containing ${}^t(1, \dots, 1)$ and a basis of $\ker M$, M will be represented by a diagonal matrix. The same holds for $A = bM + (a-b)I_n$.

6.5 Appendix

6.5.1 Eigenvectors of AB and of BA

Let A be a matrix of size $m \times n$ and B a matrix of size $n \times m$. Then the product AB is a well-defined square matrix of size $m \times m$ while BA is a well-defined square matrix of size $n \times n$. We have the following result.

Proposition. *The nonzero eigenvalues of AB and BA are the same.*

PROOF. Let $\mathbf{v} \neq 0$ be an eigenvector of AB, say $AB\mathbf{v} = \lambda\mathbf{v}$ where $\lambda \neq 0$. Then $(BA)B\mathbf{v} = B(AB)\mathbf{v} = B(\lambda\mathbf{v}) = \lambda B\mathbf{v}$. But $B\mathbf{v} \neq 0$ since $AB\mathbf{v} = \lambda\mathbf{v} \neq 0$, and hence $B\mathbf{v}$ is an eigenvector of BA relatively to the same eigenvalue λ. ■

The preceding proof shows that for $\lambda \neq 0$, $V_\lambda(AB) \subset V_\lambda(BA)$, the converse inclusion being true by symmetry. Hence

$$V_\lambda(AB) = V_\lambda(BA) \quad \text{if } \lambda \neq 0.$$

This proves the following statement.

Corollary. *The geometric multiplicities of the nonzero eigenvalues of AB and BA are the same.* ■

6.5.2 Complements on the Fibonacci Numbers

Let us simply indicate a few more properties of the Fibonacci numbers

$$\begin{aligned}
f_1 + f_2 + \cdots + f_n &= f_{n+2} - 1, \\
f_1 + f_3 + \cdots + f_{2n-1} &= f_{2n}, \\
f_1 - f_2 + \cdots + (-1)^{n+1} f_n &= (-1)^{n+1} f_{n-1} + 1, \\
f_1^2 + f_2^2 + \cdots + f_n^2 &= f_n f_{n+1}, \\
f_n^2 + f_{n+1}^2 &= f_{2n}, \\
\gcd(f_n, f_m) &= f_{\gcd(n,m)}.
\end{aligned}$$

The last formula justifies the numbering starting with $f_0 = 0$, $f_1 = 1$: With this choice, if n divides m then f_n also divides f_m, and the arithmetic properties of the sequence (f_n) are most easily formulated.

From $f_4 = (3/2) f_3$, it follows that $f_n \geqslant (3/2)^{n-2}$ for all $n \geqslant 4$. More precisely, one has

$$f_n \geqslant (1,61)^{n-2} \quad \text{for all } n \geqslant 4,$$

but $f_n < (1,62)^{n-2}$ for infinitely many values of n.

The positive root λ of $x^2 = x + 1$ is the *golden section*. By definition $\lambda = 1 + \frac{1}{\lambda}$, and iteration leads to the formula

$$\lambda = 1 + \frac{1}{\lambda} = 1 + \frac{1}{1 + \frac{1}{\lambda}} = \cdots = 1 + \frac{1}{1 + \frac{1}{1 + \frac{1}{1 + \cdots}}}.$$

Also by definition, $\lambda = \sqrt{1 + \lambda}$, and iteration now leads to

$$\lambda = \sqrt{1 + \sqrt{1 + \lambda}} = \cdots = \sqrt{1 + \sqrt{1 + \sqrt{1 + \sqrt{1 + \cdots}}}}.$$

6.6 Exercises

1. Find the eigenvectors of

$$\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad \begin{pmatrix} \frac{3}{5} & \frac{1}{5} \\ \frac{2}{5} & \frac{4}{5} \end{pmatrix}.$$

2. If $\mathbf{v}$ is an eigenvector of an operator T, and S is any invertible operator, show that $S^{-1}\mathbf{v}$ is an eigenvector of $T' = S^{-1}TS$ with the same eigenvalue.

3. If T is an operator satisfying $T^2 = -T$, what are the possible eigenvalues of T? Give a few examples of 3×3 matrices satisfying this relation.

4. Show that if a power of N vanishes, the only eigenvalue of $I \pm N$ is 1.

5. Let M denote the $n \times n$ matrix

$$M = \begin{pmatrix} 1 & 1 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix}.$$

Observe that $M^2 = nM$: What are the possible eigenvalues of M? What are the possible eigenvalues of $aI_n + bM$? (As we have seen in (6.4.2), all these matrices are diagonalizable.)

6. Determine the number of ways T_n a train of length n can be built using carriages of length 1 and 2 only. Here are the first values

$T_1 = 1$:

$T_2 = 2$:

$T_3 = 3$:

$T_4 = 5$:

7. Show that the matrices $\begin{pmatrix} 3 & 1 \\ 0 & 3 \end{pmatrix}$, and $\begin{pmatrix} 2 & -1 \\ 1 & 0 \end{pmatrix}$, have only one eigenvalue.

8. (*a*) Show that the matrices $\begin{pmatrix} a & b \\ b & d \end{pmatrix}$ have two independent eigenvectors. Give the eigenvectors and eigenvalues explicitly for the particular matrices

$$\begin{pmatrix} a & b \\ b & 0 \end{pmatrix}, \quad \begin{pmatrix} a & b \\ b & a \end{pmatrix}, \quad \begin{pmatrix} a & b \\ b & -a \end{pmatrix}.$$

In particular, what are the eigenvectors and eigenvalues of $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$. Compute the nth power A^n of $\begin{pmatrix} a & b \\ b & a \end{pmatrix}$.

(*b*) Diagonalize

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

9. The Perrin sequence is defined by

$$\begin{cases} P(1) = 0, \ P(2) = 2, \ P(3) = 3 \text{ and} \\ P(n) = P(n-2) + P(n-3) \quad (n \geqslant 4). \end{cases}$$

Prove

$$P(n+1)/P(n) \to \alpha = 1.3247\ldots \quad (n \to \infty)$$

where α is the positive root of $x^3 = x + 1$.

10. The "Tribonacci" sequence is defined by

$$\begin{cases} T(1) = 1,\ T(2) = 1,\ T(3) = 2 \text{ and} \\ T(n) = T(n-1) + T(n-2) + T(n-3) \quad (n \geqslant 4). \end{cases}$$

Prove

$$T(n+1)/T(n) \to \beta = 1.83929\ldots \quad (n \to \infty)$$

where β is the positive root of $x^3 = x^2 + x + 1$.

11. Let

$$A = \begin{pmatrix} 1 & 0 & 2 \\ 3 & -1 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 2 \\ 3 & 1 \\ 1 & 1 \end{pmatrix}.$$

Compute the eigenvalues of the square matrices AB and BA an compare them.

12. Compute all powers A^n of the matrix

$$A = \begin{pmatrix} 2 & -1 & -1 \\ 1 & 2 & 1 \\ -1 & 1 & 2 \end{pmatrix}.$$

13. Diagonalize the following matrix

$$A = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}.$$

What are the powers A^n $(n \geqslant 1)$ of A?

14. Let V denote the space of polynomials of degree less than 5. Consider the operator T in V defined by

$$T(p) = \frac{1}{x} \int_0^x p(t)\,dt.$$

Give the matrix of T in the canonical basis of V. What are the eigenvectors and eigenvalues of T?

15. Compute

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}^n \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} a_n \\ b_n \end{pmatrix}$$

for $n = 1$, 2, 3, and 4. Can you guess a formula for a_n and b_n?

16. Prove that $f_{2n+1} f_{2n-1}$ divides $f_{2n+1}^2 + f_{2n-1}^2 + 1$ for all $n \geqslant 1$. In fact, it can be proved that

> The only pairs of positive integers $a > b$ such that ab divides $a^2 + b^2 + 1$ are $a = f_{2n+1} > b = f_{2n-1}$.

Notes

Leonardo da Pisa (≈1180-1250) is also known as Fibonacci (son of Bonaccio). He is the author of the famous *Liber Abaci* (≈1202): Book of the Abacus.

Here is another milestone

Jacques P.M. Binet: C.R.Acad.Sc. Paris, N° 17 (1843) p.559-567

and a recent edition of a classical text

Nicolai N.Vorobiev: Fibonacci Numbers, Birkhäuser Verlag (2002), ISBN: 3-7643-6135-2

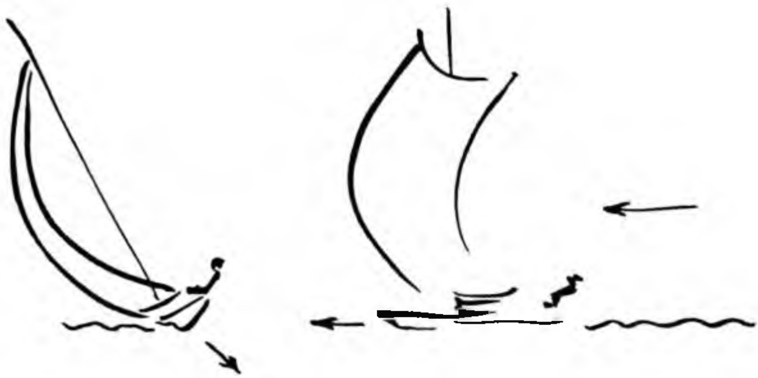
The general method for finding eigenvalues and eigenvectors based on row operations was promoted in several articles:

W. A. McWorter Jr, L.F. Meyers, Computing Eigenvalues and Eigenvectors without Determinants, Mathematics Magazine **71**, nb.2 (1998), 24-33

S. Axler: Down with determinants, Amer. Math. Monthly, **102** (1995), 139-154

Keywords for Web Search

Fibonacci phyllotaxis
Golden Ratio, golden section
Story of Phi

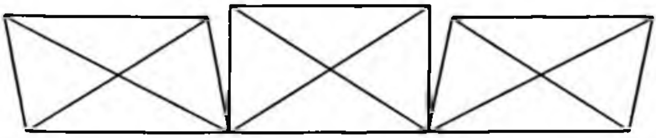


Whenever possible, use eigenvectors!

Chapter 7

Inner-Product Spaces

Here starts the ***second part*** of this book. It concerns *lengths* and *angles*. For example, the graphs of the linear functions $y = \frac{a}{b}x$ and $y = \frac{-b}{a}x$ are orthogonal. How does one recognize orthogonality of two vectors? A carpenter who has a flexible frame knows that it is a perfect rectangle when both diagonals have the same length.



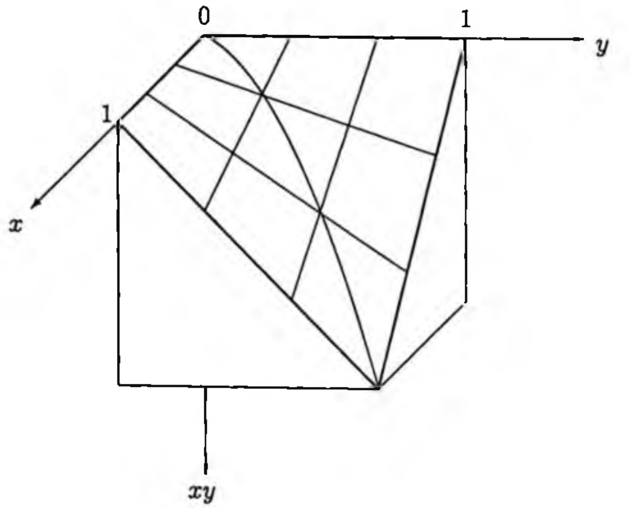
In our language, $\mathbf{u}$ and $\mathbf{v}$ are orthogonal when $\mathbf{u} + \mathbf{v}$ and $\mathbf{u} - \mathbf{v}$ have the same length. The concept of inner product—a bilinear notion—gives the general context for these considerations. Whereas the restriction to real scalars was only a simplifying assumption in the first part of this book, it is essential for inner products.

7.1 About Multiplication and Products

The multiplication of scalars is an operation formally described by a mapping

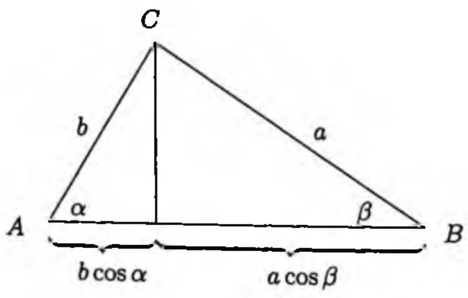
$$(x, y) \longmapsto xy : \mathbf{R}^2 \longrightarrow \mathbf{R},$$

which is not linear. However, if we fix one variable—say $y = y_0$—then the map $x \mapsto xy_0 : \mathbf{R} \to \mathbf{R}$ is linear. Hence this multiplication gives rise to two families of linear maps $x \mapsto xy_0,\ y \mapsto x_0y : \mathbf{R} \to \mathbf{R}$. For this reason, we say that the product $\mathbf{R}^2 \to \mathbf{R}$ is *bilinear*. The graph of this function is a surface containing *two families of lines*.



7.1.1 The Dot Product in Plane Geometry

Let us start with a triangle in the Euclidean plane, having sides with known positive lengths a, b, and c. Suppose that we have to determine the angles (or their cosines) of this triangle. Call α the angle at vertex A (resp. β, γ for the other angles). We know that the length of the orthogonal projection of a segment is shrunk by a factor equal to the cosine of the angle between the two directions.



A first relation between the angles α and β is found if we project the sides AC and BC onto AB

$$c = b\cos\alpha + a\cos\beta.$$

Projecting similarly onto the other sides, we are led to two further equations (which can be obtained from the first one by the use of circular permutations).

Here is the system

$$\begin{cases} b\cos\alpha & + & a\cos\beta & & & = & c \\ & & c\cos\beta & + & b\cos\gamma & = & a \\ c\cos\alpha & + & & + & a\cos\gamma & = & b. \end{cases}$$

It is linear in the three variables $x_1 = \cos\alpha$, $x_2 = \cos\beta$, and $x_3 = \cos\gamma$. Let us solve it by elimination, using row operations:

$$\left(\begin{array}{ccc|c} b & a & 0 & c \\ 0 & c & b & a \\ c & 0 & a & b \end{array}\right) \sim \left(\begin{array}{ccc|c} b & a & 0 & c \\ 0 & c & b & a \\ 0 & -\frac{c}{b}a & a & b-\frac{c}{b}c \end{array}\right) \sim \left(\begin{array}{ccc|c} b & a & 0 & c \\ 0 & c & b & a \\ 0 & 0 & 2a & b-\frac{c^2}{b}+\frac{a^2}{b} \end{array}\right).$$

We find

$$2a\cos\gamma = \frac{a^2+b^2-c^2}{b}, \quad \cos\gamma = \frac{a^2+b^2-c^2}{2ab},$$

as well as two similar expressions for the other angles. We have obtained the *law of cosines*

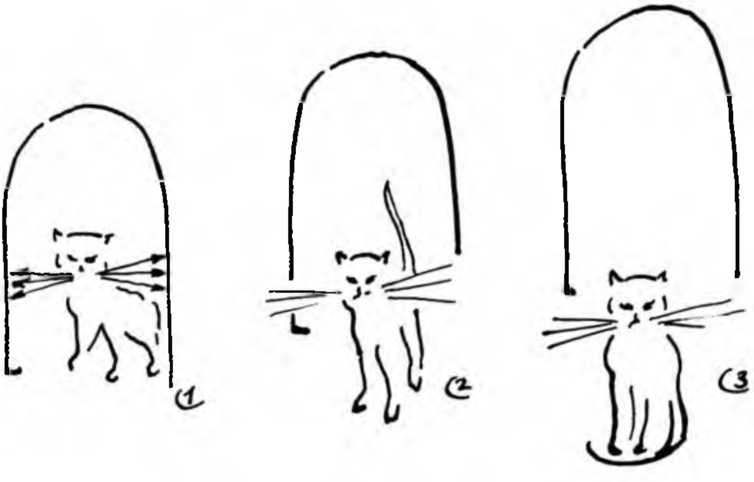
$$c^2 = a^2 + b^2 - 2ab\cos\gamma.$$

In particular, we get the *Pythagorean theorem*

$$c^2 = a^2 + b^2 \quad \Longleftrightarrow \quad \cos\gamma = 0.$$

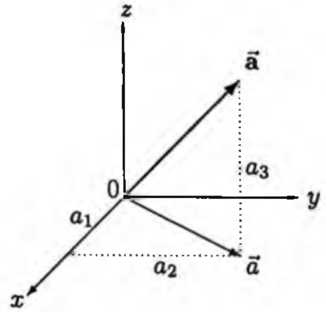
We deduce the *length* of a vector in the plane $\mathbf{R}^2$:

$$\vec{a} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \text{ has square length } \ell(\vec{a})^2 = a_1^2 + a_2^2.$$



Vectors are used for measuring lengths!

We may now find the *length* of a vector in the usual 3-space:



$$\vec{\mathbf{a}} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \text{ has square length } a^2 = \ell(\vec{\mathbf{a}})^2 = (a_1^2 + a_2^2) + a_3^2 = \sum a_i^2.$$

Let us now consider two independent vectors $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}} \in \mathbf{R}^3$, and define $\vec{\mathbf{c}} = \vec{\mathbf{b}} - \vec{\mathbf{a}}$. Since

$$a^2 = \sum a_i^2, \quad b^2 = \sum b_i^2, \quad c^2 = \sum (b_i - a_i)^2,$$

the law of cosines (in the plane generated by these two vectors) gives

$$c^2 = \sum (a_i - b_i)^2 = \sum a_i^2 + \sum b_i^2 - 2ab\cos\gamma,$$

where γ is the angle between $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$. Since

$$\sum (a_i - b_i)^2 = \sum (a_i^2 - 2a_ib_i + b_i^2),$$

we are simply left with $\sum -2a_ib_i = -2ab\cos\gamma$, and this proves

$$\sum a_ib_i = ab\cos\gamma.$$

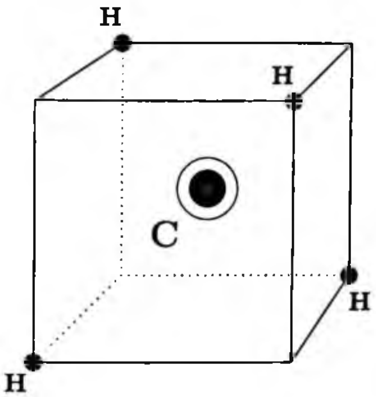
The function $F(\vec{\mathbf{a}}, \vec{\mathbf{b}}) = \sum a_ib_i$ is linear in $\vec{\mathbf{a}}$ if $\vec{\mathbf{b}}$ is fixed, symmetric in $\vec{\mathbf{a}}$ and $\vec{\mathbf{b}}$, hence also linear in $\vec{\mathbf{b}}$ if $\vec{\mathbf{a}}$ is fixed. It is a *bilinear function*. The right-hand side $ab\cos\gamma$ has a geometrical meaning, whence its interest. For example,

$$\vec{\mathbf{a}} \text{ orthogonal to } \vec{\mathbf{b}} \iff \cos\gamma = 0 \iff \sum a_ib_i = 0.$$

Application. The four hydrogen atoms in the methane molecule CH_4 are placed symmetrically so that the angles $\measuredangle HCH$ for any pair of hydrogen atoms are equal. Here is a computation of this angle. The tetrahedron having vertices

$$\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \\ -1 \end{pmatrix},$$

has for edges the diagonals of faces of a cube, hence is a *regular tetrahedron* (see picture). We place the hydrogen atoms at the vertices of this regular tetrahedron, and the carbon atom at the origin which is its center of gravity.



Let us compute the angle $\gamma = \measuredangle HCH$ formed by the two top hydrogen atoms. It is the angle between the vectors

$$\begin{pmatrix}1\\1\\1\end{pmatrix}, \quad \begin{pmatrix}-1\\-1\\1\end{pmatrix}.$$

Since these vectors have the same length $\sqrt{3}$, their dot product is

$$\sqrt{3}\sqrt{3}\cos\gamma = \begin{pmatrix}1\\1\\1\end{pmatrix} \cdot \begin{pmatrix}-1\\-1\\1\end{pmatrix} = -1 - 1 + 1 = -1,$$

whence $\cos\gamma = -\frac{1}{3}$ and γ has a value to two decimal places of 109.47 degrees.

7.1.2 The Dot Product in $\mathbf{R}^n$

Definition. *The* dot product *of two n-tuples* $\mathbf{x} = (x_i)$, $\mathbf{y} = (y_i) \in \mathbf{R}^n$ *is*

$$\mathbf{x} \cdot \mathbf{y} = \sum_{1 \leqslant i \leqslant n} x_i y_i = {}^t\mathbf{x}\,\mathbf{y} \quad (\textit{matrix product}).$$

The norm $\|\mathbf{x}\| \geqslant 0$ *of* $\mathbf{x}$, *also called* length *of* $\mathbf{x}$, *is given by*

$$\|\mathbf{x}\|^2 = \mathbf{x} \cdot \mathbf{x} = \sum_{1 \leqslant i \leqslant n} x_i^2 = {}^t\mathbf{x}\,\mathbf{x} \quad (\textit{matrix product}).$$

We expect to be able to interpret $\mathbf{x}\cdot\mathbf{y}$ as the product of $\|\mathbf{x}\|\|\mathbf{y}\|$ by the cosine of a *generalized angle* between $\mathbf{x}$ and $\mathbf{y}$. This possibility rests on the following inequalities

$$-1 \leqslant \frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\,\|\mathbf{y}\|} \leqslant +1 \quad (\mathbf{x},\mathbf{y}\in\mathbf{R}^n).$$

In the context of n-tuples, this is the *Cauchy* inequality

$$|\mathbf{x}\cdot\mathbf{y}| \leqslant \|\mathbf{x}\|\,\|\mathbf{y}\|.$$

It is a consequence of the following more precise result.

Proposition. *For* $\mathbf{x}=(x_i)$, $\mathbf{y}=(y_i)\in\mathbf{R}^n$, *we have*

$$\|\mathbf{x}\|^2\|\mathbf{y}\|^2-(\mathbf{x}\cdot\mathbf{y})^2=\sum_{i<j}(x_iy_j-x_jy_i)^2\geqslant 0.$$

PROOF. Let us simply compute the right-hand side

$$\begin{aligned}
\sum_{i<j}(x_iy_j-x_jy_i)^2 &= \tfrac{1}{2}\sum_{i\neq j}(x_iy_j-x_jy_i)^2\\
&= \tfrac{1}{2}\sum_{\text{all } i,j}(x_i^2y_j^2+x_j^2y_i^2-2x_iy_jx_jy_i)\\
&= \tfrac{1}{2}\sum_{\text{all } i,j}(x_i^2y_j^2+x_j^2y_i^2-2x_iy_ix_jy_j)\\
&= \sum_{\text{all } i,j}(x_i^2y_j^2-x_iy_ix_jy_j)\\
&= \sum_i x_i^2\sum_j y_j^2-\sum_i x_iy_i\sum_j x_jy_j\\
&= \|\mathbf{x}\|^2\|\mathbf{y}\|^2-(\mathbf{x}\cdot\mathbf{y})^2.
\end{aligned}$$

Since a sum of squares is nonnegative, the result follows. ■

For $n\geqslant 1$, there are $\frac{1}{2}n(n-1)$ pairs $1\leqslant i<j\leqslant n$. In particular when $n=3$ there are only three pairs $1\leqslant i<j\leqslant 3$, and the corresponding $x_iy_j-x_jy_i$, namely

$$x_2y_3-x_3y_2,\ x_3y_1-x_1y_3,\ x_1y_2-x_2y_1$$

may be considered as the components of a vector in $\mathbf{R}^3$, hence a definition of the *cross product* in the 3-dimensional space. This case deserves special attention, and we shall study it systematically in Sec. 10.1.2.

7.2 Abstract Inner Products and Norms

7.2.1 Definition and Examples

Let E be a real vector space. An *inner product* in E is a real-valued map F of two variables $E\times E\to\mathbf{R}$, such that:

(P1) $\mathbf{x} \mapsto F(\mathbf{x}, \mathbf{y})$ *is linear for all* $\mathbf{y} \in E$

(P2) $F(\mathbf{y}, \mathbf{x}) = F(\mathbf{x}, \mathbf{y}) \quad (\mathbf{x}, \mathbf{y} \in E)$

(P3) $F(\mathbf{x}, \mathbf{x}) > 0$ *for all* $\mathbf{x} \neq 0$.

By the symmetry ($P2$), the map $\mathbf{y} \mapsto F(\mathbf{x}, \mathbf{y})$ is also linear for each $\mathbf{x} \in E$, so that F is *bilinear*. Observe that the linearity in each variable separately implies

$$F(\mathbf{x}, 0) = F(0, \mathbf{y}) = 0 \quad (\mathbf{x}, \mathbf{y} \in E).$$

As a rule, an inner product will be denoted by $(\mathbf{x} \mid \mathbf{y})$, instead of $F(\mathbf{x}, \mathbf{y})$.

Definition. *A pair consisting of a vector space E and an inner product in E, is called* inner-product space. *In an inner-product space E, the* norm *of an element $\mathbf{x} \in E$ is defined by*

$$\|\mathbf{x}\| = (\mathbf{x} \mid \mathbf{x})^{1/2} = F(\mathbf{x}, \mathbf{x})^{1/2}.$$

A unit vector *is any $\mathbf{x} \in E$ with $\|\mathbf{x}\| = 1$. A finite-dimensional inner-product space is called a* Euclidean space.

Observe that by definition of the norm, we have

$$\|\lambda \mathbf{x}\| = |\lambda| \, \|\mathbf{x}\| \quad (\lambda \in \mathbf{R}, \mathbf{x} \in E).$$

Examples. (1) The canonical inner product in $\mathbf{R}^n$ is the dot product (Sec. 7.1.2)

$$\mathbf{x} \cdot \mathbf{y} = \textstyle\sum_{1 \leqslant i \leqslant n} x_i y_i = {}^t\mathbf{x}\, \mathbf{y}.$$

In general, if F is any inner product in $\mathbf{R}^n$, we may consider the matrix

$$G = (g_{ij}) : \quad g_{ij} = F(\mathbf{e}_i, \mathbf{e}_j) = F(\mathbf{e}_j, \mathbf{e}_i) = g_{ji}$$

(the identity matrix (δ_{ij}) for the dot product). Since F is bilinear we can expand

$$F(\Sigma_i x_i \mathbf{e}_i, \Sigma_j y_j \mathbf{e}_j) = \textstyle\sum_{ij} g_{ij} x_i y_j,$$

whence a matrix formulation for this inner product

$$F(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot G\mathbf{y} = {}^t\mathbf{x}\, G\mathbf{y}.$$

By assumption

$$\mathbf{x} \cdot G\mathbf{x} > 0 \quad \text{for all } \mathbf{x} \neq 0.$$

A symmetric matrix G having this property is said to be *positive definite*, and this property is denoted by $G \gg 0$. Starting with any *invertible* matrix A of size $n \times n$, we may consider the inner product in $\mathbf{R}^n$ defined by $F(\mathbf{x}, \mathbf{y}) = A\mathbf{x} \cdot A\mathbf{y}$. In this case

$$F(\mathbf{x}, \mathbf{y}) = {}^t(A\mathbf{x})\, A\mathbf{y} = {}^t\mathbf{x}\, {}^tA A\mathbf{y} = \mathbf{x} \cdot {}^tA A\mathbf{y},$$

so that the corresponding symmetric matrix is $G = {}^tAA \gg 0$.

(2) In the infinite-dimensional vector space $E = \mathcal{C}([0,1])$, we may define an inner product by

$$(f \mid g) = \int_0^1 f(t)g(t)\,dt.$$

More generally, we may choose a positive, continuous density $w > 0$ on $[0,1]$ and define an inner product by

$$(f \mid g)_w = \int_0^1 f(t)g(t)w(t)\,dt.$$

7.2.2 The Cauchy–Schwarz–Bunyakovskiĭ Inequality

In the context of inner-product spaces, here is a fundamental inequality.

Theorem. *Let E be an inner-product space. Then*

$$|(\mathbf{x} \mid \mathbf{y})| \leqslant \|\mathbf{x}\|\,\|\mathbf{y}\| \quad (\mathbf{x},\mathbf{y} \in E),$$

with an equality $|(\mathbf{x} \mid \mathbf{y})| = \|\mathbf{x}\|\,\|\mathbf{y}\|$ precisely when $\mathbf{x}$ and $\mathbf{y}$ are proportional.

PROOF. If $\mathbf{x} = 0$, there is nothing to prove. Let us assume $\|\mathbf{x}\| \neq 0$ from now on. By positivity of inner products

$$0 \leq (t\mathbf{x} + \mathbf{y} \mid t\mathbf{x} + \mathbf{y}) = \|\mathbf{x}\|^2\,t^2 + 2(\mathbf{x} \mid \mathbf{y})\,t + \|\mathbf{y}\|^2$$

for all scalars t. Since $\|\mathbf{x}\| \neq 0$, this is a quadratic function of t, and since it does not change sign, its discriminant is nonpositive

$$4(\mathbf{x} \mid \mathbf{y})^2 - 4\|\mathbf{x}\|^2\|\mathbf{y}\|^2 \leqslant 0.$$

Moreover, if this discriminant vanishes, there is a real root, hence a value of t in $\mathbf{R}$ for which the norm of $t\mathbf{x} + \mathbf{y}$ vanishes: $\mathbf{x}$ and $\mathbf{y}$ are proportional. ■

Comment. If $\mathbf{x} \neq 0$, taking explicitly

$$t \doteq -(\mathbf{x} \mid \mathbf{y})/\|\mathbf{x}\|^2,$$

the single inequality $(t\mathbf{x} + \mathbf{y} \mid t\mathbf{x} + \mathbf{y}) \geqslant 0$ also leads to the theorem.

Definition. *In an inner-product space, two vectors $\mathbf{x}$ and $\mathbf{y}$ are* orthogonal, *denoted by $\mathbf{x} \perp \mathbf{y}$, when their inner product vanishes: $(\mathbf{x} \mid \mathbf{y}) = 0$.*

As we have seen, the Cauchy–Schwarz inequality shows

$$-1 \leqslant \frac{(\mathbf{x} \mid \mathbf{y})}{\|\mathbf{x}\|\,\|\mathbf{y}\|} \leqslant 1$$

and this quotient $(\mathbf{x} \mid \mathbf{y})/(\|\mathbf{x}\|\,\|\mathbf{y}\|) \in [-1,1]$ may be interpreted as the cosine of an *angle*

$$\theta = \sphericalangle(\mathbf{x},\mathbf{y}) \in [0,\pi] \quad \Longleftrightarrow \quad \cos\theta = \frac{(\mathbf{x} \mid \mathbf{y})}{\|\mathbf{x}\|\,\|\mathbf{y}\|} \in [-1,1].$$

With this interpretation, the angle between two orthogonal vectors is a right angle, since its cosine vanishes.

7.2.3 The Pythagorean Theorem

In any inner-product space, we say that two vectors $\mathbf{x}$, $\mathbf{y}$ are orthogonal when $(\mathbf{x} \mid \mathbf{y}) = 0$ and denote this by $\mathbf{x} \perp \mathbf{y}$. We can now give the general version of the Pythagorean theorem.

Theorem. *In any inner-product space*

$$\mathbf{x} \perp \mathbf{y} \iff \|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2.$$

PROOF. Let us compute the square of the norm of $\mathbf{x}+\mathbf{y}$ using the bi-linearity of the inner product

$$\begin{aligned} \|\mathbf{x}+\mathbf{y}\|^2 &= (\mathbf{x}+\mathbf{y} \mid \mathbf{x}+\mathbf{y}) = (\mathbf{x} \mid \mathbf{x}+\mathbf{y}) + (\mathbf{y} \mid \mathbf{x}+\mathbf{y}) \\ &= (\mathbf{x} \mid \mathbf{x}) + (\mathbf{x} \mid \mathbf{y}) + (\mathbf{y} \mid \mathbf{x}) + (\mathbf{y} \mid \mathbf{y}) \\ &= \|\mathbf{x}\|^2 + 2(\mathbf{x} \mid \mathbf{y}) + \|\mathbf{y}\|^2. \end{aligned}$$

The conclusion follows. ■

A family $(\mathbf{x}_i)_{i\in I}$ of elements of an inner-product space E consisting of mutually orthogonal vectors $\mathbf{x}_i \perp \mathbf{x}_j$ $(i \neq j)$, is simply called *an orthogonal family* of E.

Corollary 1. *For any finite orthogonal family* $(\mathbf{x}_i)_{1\leqslant i\leqslant k}$, *we have*

$$\Big\| \sum_{1\leqslant i\leqslant k} \mathbf{x}_i \Big\|^2 = \sum_{1\leqslant i\leqslant k} \|\mathbf{x}_i\|^2.$$

PROOF. We can make an induction on the number of vectors, based on the theorem for two vectors. For the case of $n > 2$ vectors, we observe

$$\mathbf{x}_1 \perp \textstyle\sum_{i>1} \mathbf{x}_i \implies \big\|\mathbf{x}_1 + (\sum_{i>1} \mathbf{x}_i)\big\|^2 = \|\mathbf{x}_1\|^2 + \big\|\sum_{i>1} \mathbf{x}_i\big\|^2.$$

Hence the induction hypothesis furnishes the conclusion. ■

Corollary 2. *Any orthogonal family not containing the zero vector is linearly independent.*

PROOF. Let $(\mathbf{x}_i)_{1\leqslant i\leqslant k}$ be a finite family consisting of nonzero mutually orthogonal vectors $\mathbf{x}_i \perp \mathbf{x}_j$ $(i \neq j)$. If $\sum a_i\mathbf{x}_i = 0$, then

$$\textstyle\sum_i a_i^2\|\mathbf{x}_i\|^2 = \big\|\sum_i a_i\mathbf{x}_i\big\|^2 = 0$$

shows that all summands $a_i^2\|\mathbf{x}_i\|^2$ vanish. Since $\|\mathbf{x}_i\| \neq 0$, we conclude $a_i = 0$ for all indices i. ■

OTHER PROOF. Assume $\sum_i a_i\mathbf{x}_i = 0$. Then

$$0 = \big(\mathbf{x}_j \mid \textstyle\sum_i a_i\mathbf{x}_i\big) = \sum_i a_i(\mathbf{x}_j \mid \mathbf{x}_i) = a_j\|\mathbf{x}_j\|^2,$$

hence $a_j = 0$ (since $\|\mathbf{x}_j\| \neq 0$ by assumption), for all indices j. ■

7.2.4 More Identities

Let us come back to the identities

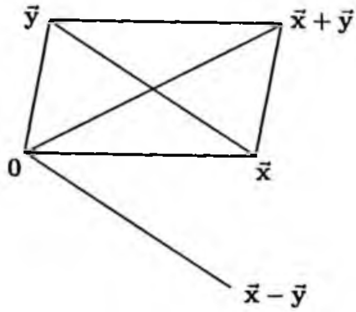
$$\begin{aligned}\|\mathbf{x}+\mathbf{y}\|^2 &= \|\mathbf{x}\|^2 + 2(\mathbf{x} \mid \mathbf{y}) + \|\mathbf{y}\|^2,\\ \|\mathbf{x}-\mathbf{y}\|^2 &= \|\mathbf{x}\|^2 - 2(\mathbf{x} \mid \mathbf{y}) + \|\mathbf{y}\|^2.\end{aligned}$$

(1) If we add them, we find

$$\|\mathbf{x}+\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{y}\|^2 = 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2.$$

This is the *parallelogram equality*:

> *The sum of the squares of the sides is equal to the sum of the squares of the diagonals.*



Equivalently

$$\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 = \frac{\|\mathbf{x}+\mathbf{y}\|^2 + \|\mathbf{x}-\mathbf{y}\|^2}{2},$$

shows that the sum $\|\mathbf{x}\|^2+\|\mathbf{y}\|^2$ is the mean value of the squares of the diagonals. The condition $\|\mathbf{x}+\mathbf{y}\| = \|\mathbf{x}-\mathbf{y}\|$ characterizes rectangles among parallelograms.

(2) If we subtract them, we find

$$\|\mathbf{x}+\mathbf{y}\|^2 - \|\mathbf{x}-\mathbf{y}\|^2 = 4(\mathbf{x} \mid \mathbf{y}),$$

which shows that the inner product of two vectors can be computed from their norms. This is often referred to as the *polarization identity*.

***Proposition** (**Minkowski Inequality**). In any inner-product space, we have*

$$\|\mathbf{x}+\mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|.$$

PROOF. Let us start once more from the square of the norm of a sum

$$\|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + 2(\mathbf{x} \mid \mathbf{y}) + \|\mathbf{y}\|^2.$$

The Cauchy–Schwarz inequality leads to

$$\|\mathbf{x}+\mathbf{y}\|^2 \leqslant \|\mathbf{x}\|^2 + 2\|\mathbf{x}\|\|\mathbf{y}\| + \|\mathbf{y}\|^2 = \left(\|\mathbf{x}\| + \|\mathbf{y}\|\right)^2,$$

hence the announced inequality. ∎

Definition. *A* norm *in a real vector space E is any map*

$$\begin{aligned} E &\longrightarrow \mathbf{R} \\ \mathbf{x} &\longmapsto \|\mathbf{x}\|, \end{aligned}$$

satisfying the following axioms:

$$\begin{aligned} &\text{(N1)} \quad \|\mathbf{x}\| > 0 \ \textit{ for all } \mathbf{x} \neq 0 \\ &\text{(N2)} \quad \|a\mathbf{x}\| = |a|\|\mathbf{x}\| \quad (a \text{ scalar}) \\ &\text{(N3)} \quad \|\mathbf{x}+\mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|. \end{aligned}$$

A normed space *is a pair* $(E, \|\cdot\|)$ *consisting of a (real) vector space E and a norm* $\|\cdot\| : E \to \mathbf{R}$.

Any inner-product space is a normed space: (N1) and (N2) hold trivially, while (N3) is the Minkowski inequality, which has just been proved. Let us give more properties of the norm in any inner-product space.

Proposition. *When the norm is deduced from an inner product, we have*

$$\|\mathbf{u}\| = \|\mathbf{v}\| \quad \Longrightarrow \quad \|a\mathbf{u} + b\mathbf{v}\| = \|b\mathbf{u} + a\mathbf{v}\| \quad (a, b \in \mathbf{R}).$$

For any pair of nonzero vectors $\mathbf{u}$, $\mathbf{v}$, *we also have*

$$\left\| \frac{\mathbf{u}}{\|\mathbf{u}\|^2} - \frac{\mathbf{v}}{\|\mathbf{v}\|^2} \right\| = \frac{\|\mathbf{u}-\mathbf{v}\|}{\|\mathbf{u}\|\,\|\mathbf{v}\|}.$$

PROOF. Simply compute the square of the norm using the inner product:

$$\|a\mathbf{u}+b\mathbf{v}\|^2 = (a\mathbf{u}+b\mathbf{v} \mid a\mathbf{u}+b\mathbf{v}) = a^2\|\mathbf{u}\|^2 + 2ab(\mathbf{u} \mid \mathbf{v}) + b^2\|\mathbf{v}\|^2.$$

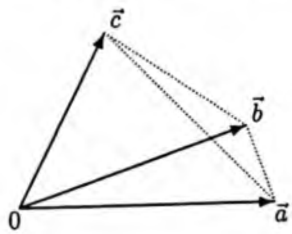
When $\|\mathbf{u}\| = \|\mathbf{v}\|$, this expression is symmetric in $\mathbf{u}$ and $\mathbf{v}$, whence the first assertion. For the second assertion, let us denote by $u = \|\mathbf{u}\|$, $\mathbf{u}_1 = \mathbf{u}/u$, $\mathbf{u}' = \mathbf{u}_1/u = \mathbf{u}/u^2$, and similarly for $\mathbf{v}$. The first part may be applied to the unit vectors $\mathbf{u}_1$ and $\mathbf{v}_1$

$$\|\mathbf{u}'-\mathbf{v}'\| = \left\|\tfrac{1}{u}\mathbf{u}_1 - \tfrac{1}{v}\mathbf{v}_1\right\| = \left\| -\tfrac{1}{v}\mathbf{u}_1 + \tfrac{1}{u}\mathbf{v}_1\right\| = \frac{\|-\mathbf{u}+\mathbf{v}\|}{uv},$$

which is the announced result. ■

Corollary. *Let $\vec{a}$, $\vec{b}$, and $\vec{c}$ be three nonzero vectors in the Euclidean plane* $\mathbf{R}^2$. *Then*

$$\|\vec{b}\|\,\|\vec{a}-\vec{c}\| \leqslant \|\vec{c}\|\,\|\vec{a}-\vec{b}\| + \|\vec{a}\|\,\|\vec{b}-\vec{c}\|.$$



PROOF. As in the proof of the proposition, let us use the notation $a = \|\vec{a}\|$, $\vec{a}' = \vec{a}/a^2$, and similarly for $\vec{b}$ and $\vec{c}$. The triangle inequality gives

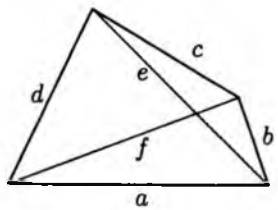
$$\|\vec{a}' - \vec{c}'\| \leqslant \|\vec{a}' - \vec{b}'\| + \|\vec{b}' - \vec{c}'\|,$$

with equality precisely when the extremities of $\vec{a}'$, $\vec{b}'$, and $\vec{c}'$ are on a straight line. Using the proposition, we deduce

$$\frac{\|\vec{a} - \vec{c}\|}{ac} \leqslant \frac{\|\vec{a} - \vec{b}\|}{ab} + \frac{\|\vec{b} - \vec{c}\|}{bc}.$$

Multiplying this inequality by abc, we get the announced result. ∎

With another notation, the result is easier to remember:



$$ef \leqslant ac + bd.$$

This is the *Ptolemy inequality*. The equality $ef = ac + bd$ holds precisely when the vertices of the quadrilateral lie on a circle: *Ptolemy's theorem*. Indeed in this case, it is well known that the inversion

$$\vec{x} \longmapsto \vec{x}' = \vec{x}/x^2 = \vec{x}_1/x,$$

transforms this circle into a straight line, and the triangle inequality becomes an equality.

Theorem. *Let E be an inner-product space, and $(\mathbf{a}_i)_{1\leqslant i\leqslant n}$ a finite family in E. Consider the $n \times n$ symmetric matrix $A = (a_{ij})$ having entries $a_{ij} = \|\mathbf{a}_i - \mathbf{a}_j\|^2$. Then for any $\mathbf{x} = (x_i) \in \mathbf{R}^n$ with $\sum_i x_i = 0$ we have*

$$(\mathbf{x} \mid A\mathbf{x}) = \sum_{1\leqslant i,j\leqslant n} x_i x_j \|\mathbf{a}_i - \mathbf{a}_j\|^2 \leqslant 0.$$

PROOF. Since

$$\|\mathbf{a}_i - \mathbf{a}_j\|^2 = (\mathbf{a}_i - \mathbf{a}_j \mid \mathbf{a}_i - \mathbf{a}_j) = \|\mathbf{a}_i\|^2 - 2(\mathbf{a}_i \mid \mathbf{a}_j) + \|\mathbf{a}_j\|^2,$$

the inner product (in $\mathbf{R}^n$) $(\mathbf{x} \mid A\mathbf{x})$ is the sum of three terms

$$\sum_{1\leqslant i,j\leqslant n} x_i x_j \|\mathbf{a}_i\|^2 = \textstyle\sum_i x_i \|\mathbf{a}_i\|^2 \underbrace{\sum_j x_j}_{=0} = 0,$$

$$\sum_{1\leqslant i,j\leqslant n} x_i x_j \|\mathbf{a}_j\|^2 = \underbrace{\textstyle\sum_i x_i}_{=0} \sum_j x_j \|\mathbf{a}_j\|^2 = 0,$$

and

$$-2\sum_{1\leqslant i,j\leqslant n} x_i x_j (\mathbf{a}_i \mid \mathbf{a}_j) = -2\Big(\textstyle\sum_i x_i \mathbf{a}_i \,\Big|\, \sum_j x_j \mathbf{a}_j\Big) = -2\Big\|\textstyle\sum_i x_i \mathbf{a}_i\Big\|^2 \leqslant 0.$$

The announced result follows. ■

7.3 Orthonormal Bases

An orthonormal family in an inner-product space E is simply an orthogonal family consisting of unit (namely normalized) elements.

Definition. *An* orthonormal basis *of an inner-product space E is a basis $(\mathbf{e}_i)_{i\in I}$ consisting of mutually orthogonal unit vectors, namely such that*

$$\|\mathbf{e}_i\| = 1, \quad \mathbf{e}_i \perp \mathbf{e}_j \quad (i \neq j \in I).$$

We may abbreviate these conditions with the Kronecker symbol:

$$(\mathbf{e}_i \mid \mathbf{e}_j) = \delta_{ij} \quad (i, j \in I).$$

For example, the canonical basis $(\mathbf{e}_i)_{1\leqslant i\leqslant n}$ of $\mathbf{R}^n$ is an orthonormal basis. Observe that if $\mathbf{x} = (x_i) \in \mathbf{R}^n$, then $x_i = \mathbf{x}\cdot\mathbf{e}_i$. Here is the generalization to any inner-product space.

Theorem. *Let $(\mathbf{e}_i)_{i\in I}$ be an orthonormal basis of an inner-product space E. Then the components of an element $\mathbf{x} = \sum_{i\in I} x_i \mathbf{e}_i \in E$ are given by the inner products*

$$x_i = (\mathbf{x} \mid \mathbf{e}_i) \quad (i \in I).$$

PROOF. If $\mathbf{x} = \sum_{i\in I} x_i \mathbf{e}_i$ (a finite linear combination), take the inner product with any basis vector $\mathbf{e}_j$:

$$(\mathbf{x} \mid \mathbf{e}_j) = \Big(\sum_{i\in I} x_i \mathbf{e}_i \,\Big|\, \mathbf{e}_j\Big) = \sum_{i\in I} x_i \underbrace{(\mathbf{e}_i \mid \mathbf{e}_j)}_{=\delta_{ij}} = x_j. \qquad ■$$

7.3.1 Euclidean Spaces

A finite-dimensional inner-product space will be called a *Euclidean space*. In such a space, it is easy to prove that there is an orthonormal basis.

Theorem 1. *Any Euclidean space E has an orthonormal basis $(\mathbf{e}_i)_{1\leqslant i\leqslant n}$.*

PROOF. The theorem is trivial if $\dim E = 1$, and we prove it by induction on the dimension n of E. If $n > 1$, we can find a normed vector $\mathbf{e}_0$ in E: Take any nonzero $\mathbf{x} \in E$ and put $\mathbf{e}_0 = \frac{1}{\|\mathbf{x}\|}\mathbf{x}$. Consider the vector subspace

$$\begin{aligned} V &= \{\mathbf{y} \in E : (\mathbf{y} \mid \mathbf{e}_0) = 0\} \\ &= \ker(\mathbf{y} \mapsto (\mathbf{y} \mid \mathbf{e}_0)) \subset E, \end{aligned}$$

which does not contain $\mathbf{e}_0$, hence $V \neq E$: $\dim V < n = \dim E$. By induction hypothesis, we can find an orthonormal basis $\mathbf{e}_1, \ldots, \mathbf{e}_m$ of V. I claim that $\mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_m$ is a basis of E. Since any orthonormal family is linearly independent, it is sufficient to prove that it generates E. But if $\mathbf{x} \in E$, observe that $\mathbf{x} - (\mathbf{x} \mid \mathbf{e}_0)\mathbf{e}_0 \in V$

$$(\mathbf{x} - (\mathbf{x} \mid \mathbf{e}_0)\mathbf{e}_0 \mid \mathbf{e}_0) = (\mathbf{x} \mid \mathbf{e}_0) - (\mathbf{x} \mid \mathbf{e}_0)(\mathbf{e}_0 \mid \mathbf{e}_0) = 0.$$

Hence

$$\begin{aligned} \mathbf{x} - (\mathbf{x} \mid \mathbf{e}_0)\mathbf{e}_0 &= \sum_{1\leqslant i\leqslant m} x_i \mathbf{e}_i \\ \mathbf{x} &= (\mathbf{x} \mid \mathbf{e}_0)\mathbf{e}_0 + \sum_{1\leqslant i\leqslant m} x_i \mathbf{e}_i. \end{aligned}$$

This proves the claim. Moreover $n = \dim E = m + 1$ so that $m = n - 1$. ■

As an application, let us determine all inner products in $\mathbf{R}^n$, namely all real-valued functions F of two variables in $\mathbf{R}^n$ such that:

$$\begin{aligned} F(\mathbf{x}, \mathbf{y}) &= (\mathbf{x} \mid \mathbf{y})_F \quad \text{is bilinear} \\ F(\mathbf{x}, \mathbf{x}) &= \|\mathbf{x}\|_F^2 > 0 \quad \text{if } \mathbf{x} \neq 0. \end{aligned}$$

Take an orthonormal basis $(\varepsilon_i)_{1\leqslant i\leqslant n}$ of $\mathbf{R}^n$ with respect to this inner product, say

$$\varepsilon_j = \begin{pmatrix} s_{1j} \\ \vdots \\ s_{nj} \end{pmatrix} = \sum_{1\leqslant i\leqslant n} s_{ij}\mathbf{e}_i,$$

(where $(\mathbf{e}_i)$ denotes the canonical basis of $\mathbf{R}^n$). In this basis, we have

$$\mathbf{e}_j = \sum a_{ij}\varepsilon_i,$$

so that

$$g_{ij} = (\mathbf{e}_i \mid \mathbf{e}_j)_F = \sum_{k,\ell} a_{ki}a_{\ell j}\delta_{k\ell} = \sum_k a'_{ik}a_{kj} \quad (a'_{ik} = a_{ki}).$$

These identities show that the matrix $G = (g_{ij})$ is equal to the matrix product ${}^tAA$ where $A = (a_{ij})$. Hence we have

$$\begin{aligned} (\mathbf{x} \mid \mathbf{y})_F &= (\Sigma_i x_i \mathbf{e}_i \mid \Sigma_j y_j \mathbf{e}_j) = \textstyle\sum_{i,j} x_i g_{ij} y_j \\ &= {}^t\mathbf{x} G \mathbf{y} = {}^t\mathbf{x}\, {}^tAA\mathbf{y} = A\mathbf{x} \cdot A\mathbf{y}. \end{aligned}$$

This proves that any inner product F in $\mathbf{R}^n$ is obtained from the dot product by means of a linear change of coordinates, according to

$$(\mathbf{x} \mid \mathbf{y})_F = {}^t(A\mathbf{x})(A\mathbf{y}) = A\mathbf{x} \cdot A\mathbf{y} \quad (\mathbf{x}, \mathbf{y} \in \mathbf{R}^n),$$

for some invertible matrix A. Remembering the first example of Sec. 7.2.1, we have obtained the following result.

Theorem 2. There is a one-to-one correspondence between inner products F in $\mathbf{R}^n$ and symmetric, positive-definite matrices $G \gg 0$ of size $n \times n$: It is given by

$$(\mathbf{x} \mid \mathbf{y})_F = {}^t\mathbf{x}\, G\, \mathbf{y}.$$

Each positive-definite matrix $G \gg 0$ can be written (in general in several ways) $G = {}^tAA$ for some invertible matrix A, so that

$$(\mathbf{x} \mid \mathbf{y})_F = A\mathbf{x} \cdot A\mathbf{y} \quad (\mathbf{x}, \mathbf{y} \in \mathbf{R}^n).$$

■

7.3.2 The Best Approximation Theorem

Theorem 1. Let E be an inner-product space, V a subspace of E, and $\mathbf{x} \in E$. The following conditions for an element $\mathbf{y} \in V$ are equivalent:

(*i*) $\mathbf{x} - \mathbf{y} \perp \mathbf{v}$ *for all* $\mathbf{v} \in V$

(*ii*) $\|\mathbf{x} - \mathbf{y}\| = \min_{\mathbf{v} \in V} \|\mathbf{x} - \mathbf{v}\|$.

There is at most one $\mathbf{y} \in V$ *satisfying them.*

PROOF. (*i*) $\Rightarrow$ (*ii*) If $\mathbf{x} - \mathbf{y} \perp V$, then for each $\mathbf{v} \in V$, we can write

$$\begin{aligned} \|\mathbf{x} - \mathbf{v}\|^2 &= \|(\mathbf{x} - \mathbf{y}) + (\mathbf{y} - \mathbf{v})\|^2 \\ &= \|\mathbf{x} - \mathbf{y}\|^2 + \|\mathbf{y} - \mathbf{v}\|^2 \\ &\geqslant \|\mathbf{x} - \mathbf{y}\|^2 \qquad\qquad (\mathbf{v} \in V), \end{aligned}$$

by the Pythagorean theorem. Hence when $\mathbf{v}$ varies in V, $\|\mathbf{x} - \mathbf{v}\|$ is minimal when $\mathbf{v} = \mathbf{y}$. As we now see, this minimum is only reached when $\|\mathbf{y} - \mathbf{v}\|^2 = 0$, hence when $\mathbf{y} = \mathbf{v}$. This gives the uniqueness part of the statement.
(*ii*) $\Rightarrow$ (*i*) For any $\mathbf{v} \in V$, the scalar function

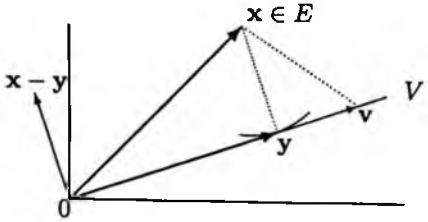
$$t \longmapsto f(t) = \|\mathbf{x} - (\mathbf{y} + t\mathbf{v})\|^2 = \big((\mathbf{x} - \mathbf{y}) - t\mathbf{v} \mid (\mathbf{x} - \mathbf{y}) - t\mathbf{v}\big)$$

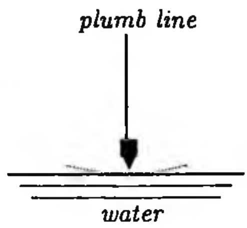
has a minimum for $t = 0$. But the equalities

$$\begin{aligned} f(t) &= \|\mathbf{x} - \mathbf{y}\|^2 - 2t(\mathbf{x} - \mathbf{y} \mid \mathbf{v}) + t^2\|\mathbf{v}\|^2, \\ f'(t) &= -2(\mathbf{x} - \mathbf{y} \mid \mathbf{v}) + 2t\|\mathbf{v}\|^2, \\ f'(0) &= -2(\mathbf{x} - \mathbf{y} \mid \mathbf{v}) = 0 \end{aligned}$$

show that $\mathbf{x} - \mathbf{y} \perp \mathbf{v}$. ■

Here are pictures illustrating the orthogonality of the best approximation.





***Theorem 2** (**Best Approximation**). Let E be an inner-product space, V the subspace generated by a finite orthonormal system $(\mathbf{e}_1, \dots \mathbf{e}_m)$ in E. Then for all $\mathbf{x} \in E$ there is a unique linear combination $\mathbf{y} = \sum_{1 \leqslant i \leqslant m} x_i \mathbf{e}_i \in V$ for which*

$$\|\mathbf{x} - \mathbf{y}\| = \min_{\mathbf{v} \in V} \|\mathbf{x} - \mathbf{v}\|.$$

This linear combination has the components $x_i = (\mathbf{x} \mid \mathbf{e}_i)$ $(1 \leqslant i \leqslant m)$.

PROOF. Take $\mathbf{y} = \sum_{1 \leqslant i \leqslant m} y_i \mathbf{e}_i$. Then

$$\begin{aligned}(\mathbf{x} - \mathbf{y} \mid \mathbf{e}_j) &= \left(\mathbf{x} - \textstyle\sum_i y_i \mathbf{e}_i \mid \mathbf{e}_j\right). \\ &= (\mathbf{x} \mid \mathbf{e}_j) - \textstyle\sum_i y_i (\mathbf{e}_i \mid \mathbf{e}_j) \\ &= (\mathbf{x} \mid \mathbf{e}_j) - \textstyle\sum_i y_i \delta_{ij} \\ &= (\mathbf{x} \mid \mathbf{e}_j) - y_j,\end{aligned}$$

so that $\mathbf{x} - \mathbf{y}$ is orthogonal to all $\mathbf{e}_j$ $(1 \leqslant j \leqslant m)$ precisely when the components y_j of $\mathbf{y}$ are given by the inner products $(\mathbf{x} \mid \mathbf{e}_j)$. Hence Theorem 2 follows from Theorem 1. ■

For any finite orthonormal family $(\mathbf{e}_j)_{j \in J}$, we can use the Pythagorean theorem to compute the norm of $\mathbf{x}$

$$\begin{aligned}\|\mathbf{x}\|^2 &= \Big\| \Big(\mathbf{x} - \sum_{j \in J} x_j \mathbf{e}_j\Big) + \sum_{j \in J} x_j \mathbf{e}_j \Big\|^2 \\ &= \Big\|\mathbf{x} - \sum_{j \in J} x_j \mathbf{e}_j\Big\|^2 + \Big\|\sum_{j \in J} x_j \mathbf{e}_j\Big\|^2,\end{aligned}$$

where

$$\Big\|\sum_{j \in J} x_j \mathbf{e}_j\Big\|^2 = \sum_{j \in J} \|x_j \mathbf{e}_j\|^2 = \sum_{j \in J} x_j^2.$$

We may conclude that $\sum_{j \in J} x_j^2 \leqslant \|\mathbf{x}\|^2$ with an equality only if $\mathbf{x}$ is a linear combination of the vectors $\mathbf{e}_j$ $(j \in J)$.

***Corollary 1** (Bessel inequality). Let $\mathbf{x} \in E$. Then, for all finite orthonormal families $(\mathbf{e}_j)_{j\in J}$, we have*

$$\sum_{j\in J}(\mathbf{x} \mid \mathbf{e}_j)^2 \leqslant \|\mathbf{x}\|^2. \qquad \blacksquare$$

Remark. The Cauchy–Schwarz inequality is a simple consequence of the Bessel inequality. If $\mathbf{y} \neq 0$, consider the orthonormal system consisting in a single vector $\mathbf{e} = \mathbf{e}_1 = \mathbf{y}/\|\mathbf{y}\|$. In this case, the Bessel inequality reads

$$(\mathbf{x} \mid \mathbf{e})^2 \leqslant \|\mathbf{x}\|^2.$$

Taking the positive square root, we deduce $|(\mathbf{x} \mid \mathbf{e})| \leqslant \|\mathbf{x}\|$, namely

$$|(\mathbf{x} \mid \mathbf{y})|/\|\mathbf{y}\| \leqslant \|\mathbf{x}\|.$$

***Corollary 2.** If E is an inner-product space and V a finite-dimensional subspace of E, then there is an orthogonal projector $P : E \to E$ having image $P(E) = V$. This projector associates to any $\mathbf{x} \in E$ the element $\mathbf{y} = P(\mathbf{x}) \in V$ which is closest to it.*

PROOF. It is enough to use the fact that V possesses an orthonormal basis $(\mathbf{e}_i)_{1\leqslant i\leqslant m}$ to be able to apply the best approximation theorem. This projector is explicitly given by

$$\mathbf{x} \longmapsto P(\mathbf{x}) = \sum_{1\leqslant i\leqslant m} (\mathbf{x} \mid \mathbf{e}_i)\, \mathbf{e}_i,$$

namely the sum of the (rank one) orthogonal projectors $P_i : \mathbf{x} \mapsto (\mathbf{x} \mid \mathbf{e}_i)\mathbf{e}_i$. $\blacksquare$

These fundamental results have numerous and important applications.

7.3.3 First Application: Periodic Functions

Let us consider the space E consisting of the continuous functions $f : [0,1] \to \mathbf{R}$ with the inner product defined by

$$(f \mid g) = \int_0^1 f(t)g(t)\,dt.$$

In particular,

$$(f \mid 1) = \int_0^1 f(t)\,dt = \text{ mean value (or average) of } f.$$

Let $\mathbf{e}_k \in E$ denote the particular functions $\mathbf{e}_k(t) = \cos 2\pi kt$ $(k \geqslant 0)$. It is well known that

$$(\mathbf{e}_k \mid \mathbf{e}_\ell) = 0 \quad \text{when } k \neq \ell,$$

while

$$2\|\mathbf{e}_k\|^2 = 2\int_0^1 \cos^2 2\pi kt\,dt = 1 \quad \text{for } k \geqslant 1.$$

Hence the functions

$$1, \quad \sqrt{2}\,\mathbf{e}_k \quad (k \geqslant 1)$$

form an orthonormal system in E. The best approximation of a function $f \in E$ by a finite linear combination $\sum_{0\leqslant k\leqslant N} a_k\mathbf{e}_k$ (N fixed) has the coefficients

$$a_0 = (f \mid 1), \quad a_k = 2(f \mid \mathbf{e}_k) \quad (1 \leqslant k \leqslant N).$$

Similar considerations hold for the orthogonal system consisting of the functions

$$\mathbf{f}_k(t) = \sin 2\pi kt \quad (k \geqslant 1).$$

Comment. The space of continuous periodic functions f with period 1

$$f(t+1) = f(t) \quad (t \in \mathbf{R}),$$

is also an inner-product space with respect to the previous definition

$$(f \mid g) = \int_0^1 f(t)g(t)\,dt.$$

Best approximations of a given periodic function f by trigonometrical polynomials of the form

$$a_0 + \sum_{1\leqslant k\leqslant N} (a_k \cos 2\pi kt + b_k \sin 2\pi kt)$$

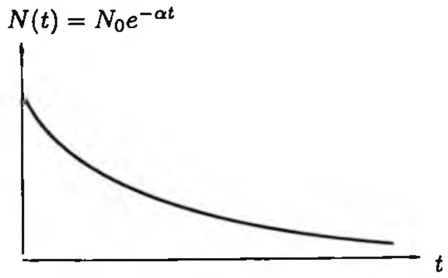
(where $N \geqslant 1$ is a fixed integer), are typically considered in the theory of Fourier series. The study of their convergence when $N \to \infty$ is not a topic tackled by linear algebra.

7.3.4 Second Application: Least Squares Method

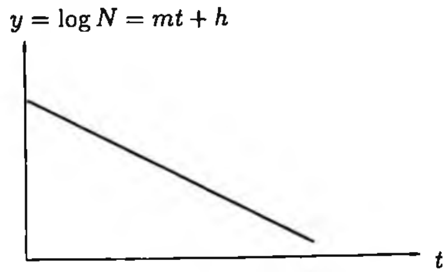
The radioactivity of a pure chemical element decreases exponentially with time: The half-life depends on the substance under consideration. If we are to identify the type of source producing such an activity, we try to determine its half-life.

Theoretically, the intensity of radioactivity is described by a function $N(t) = Ce^{-\alpha t}$ where the constants $C = N(0)$ (initial intensity) and α (type of substance) have to be determined.

Several measurements, made at different times, lead to experimental data from which the values of C and α have to be deduced. Ideally, the measurements at times t_i should detect intensities $N(t_i) = Ce^{-\alpha t_i}$.



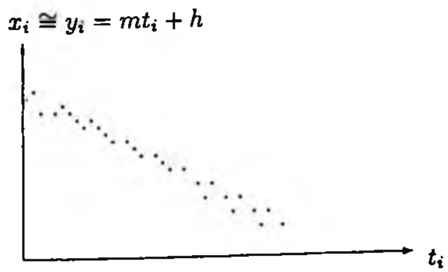
Let us adopt a logarithmic scale: $y = \log N(t) = mt + h$ ($h = \log C$, $m = -\alpha$).



In fact, the measurements produce a *cloud of points* $P_i = (t_i, x_i)$ and we have to determine a *best fit* straight line for these data. Since we are looking for a straight line, the unknown variables of the problem are m and h. The conditions linking these two variables are the equations

$$mt_i + h = x_i \quad (1 \leqslant i \leqslant n).$$

This is a linear system of n equations in 2 variables. In general, this system is incompatible and the situation seems hopeless!



The only reasonable hope is to find an *approximate* solution. Instead of

$$mt_i + h - x_i = 0 \quad (1 \leqslant i \leqslant n),$$

we introduce error terms

$$mt_i + h - x_i = \varepsilon_i \quad (1 \leqslant i \leqslant n).$$

Then m and h have to be determined in order to minimize these errors in a sensible sense. Asking for a minimal sum $\sum \varepsilon_i$ would allow one large individual ε_i to be compensated by another large opposite ε_j in the sum, hence this is not a satisfactory condition. A better idea would be to require $\sum |\varepsilon_i|$ to be minimal. Unfortunately, this condition is a challenging one, difficult to use and still open for research. The next idea is to minimize $\sum \varepsilon_i^2$: This is good enough, has a geometrical meaning and leads to an easy answer! It is known since the nineteenth century as the *least squares approximation method*.

Let us write the given system in vector form

$$m\mathbf{t} + h\mathbf{1} \cong \mathbf{x}$$

where

$$\mathbf{t} = (t_i)_{1\leqslant i\leqslant n}, \quad \mathbf{1} = (1)_{1\leqslant i\leqslant n}, \quad \mathbf{x} = (x_i)_{1\leqslant i\leqslant n} \in \mathbf{R}^n.$$

The best fit is determined by the choice of m and h so that the linear combination

$$\mathbf{y} = m\mathbf{t} + h\mathbf{1} \cong \mathbf{x}$$

is closest to the measured vector $\mathbf{x}$. This means that $\|\mathbf{y}-\mathbf{x}\|$ has to be minimized. Equivalently $\|\mathbf{y} - \mathbf{x}\|^2 = \sum \varepsilon_i^2$ has to be minimized. This best approximation is furnished by taking for $\mathbf{y}$ the *orthogonal projection* of $\mathbf{x}$ in the subspace V generated by $\mathbf{t}$ and $\mathbf{1}$. This orthogonal projection is characterized by

$$\mathbf{x} - \mathbf{y} = \mathbf{x} - m\mathbf{t} - h\mathbf{1} \quad \perp \quad \mathbf{t} \text{ and } \mathbf{1}.$$

In this way, we obtain two equations for the variables m and h:

$$\begin{cases} (\mathbf{x} \mid \mathbf{t}) - m(\mathbf{t} \mid \mathbf{t}) - h(\mathbf{1} \mid \mathbf{t}) = 0 \\ (\mathbf{x} \mid \mathbf{1}) - m(\mathbf{t} \mid \mathbf{1}) - h(\mathbf{1} \mid \mathbf{1}) = 0. \end{cases}$$

This is a linear system

$$\begin{cases} (\mathbf{t} \mid \mathbf{t})\, m + (\mathbf{1} \mid \mathbf{t})\, h = (\mathbf{x} \mid \mathbf{t}) \\ (\mathbf{t} \mid \mathbf{1})\, m + (\mathbf{1} \mid \mathbf{1})\, h = (\mathbf{x} \mid \mathbf{1}). \end{cases}$$

Incidentally, observe that if we take at least two distinct measure times t_i, the vectors $\mathbf{t}$ and $\mathbf{1}$ are independent, and the subspace $V = \mathcal{L}(\mathbf{t}, \mathbf{1})$ has dimension 2. In the Cauchy–Schwarz inequality, we have a strict inequality

$$|(\mathbf{t} \mid \mathbf{1})| < \|\mathbf{t}\|\,\|\mathbf{1}\|.$$

We shall use this in the form

$$\|\mathbf{t}\|^2\|\mathbf{1}\|^2 - |(\mathbf{t} \mid \mathbf{1})|^2 > 0.$$

In particular, this shows that the above 2×2 system has maximal rank 2. Its solution is:

$$m = \frac{(\mathbf{x} \mid \mathbf{t})(\mathbf{1} \mid \mathbf{1}) - (\mathbf{x} \mid \mathbf{1})(\mathbf{1} \mid \mathbf{t})}{(\mathbf{t} \mid \mathbf{t})(\mathbf{1} \mid \mathbf{1}) - (\mathbf{t} \mid \mathbf{1})(\mathbf{1} \mid \mathbf{t})}$$

$$h = \frac{(\mathbf{t} \mid \mathbf{t})(\mathbf{x} \mid \mathbf{1}) - (\mathbf{t} \mid \mathbf{1})(\mathbf{x} \mid \mathbf{t})}{(\mathbf{t} \mid \mathbf{t})(\mathbf{1} \mid \mathbf{1}) - (\mathbf{t} \mid \mathbf{1})(\mathbf{1} \mid \mathbf{t})}.$$

As $(\mathbf{1} \mid \mathbf{1}) = n$ (number of measurements), we can divide both numerator and denominator of these expressions by n^2 and use the notation $\langle a_i \rangle = \frac{1}{n} \sum a_i$ for averages:

$$m = \frac{\langle x_i t_i \rangle - \langle x_i \rangle \langle t_i \rangle}{\langle t_i^2 \rangle - \langle t_i \rangle^2}$$

$$h = \frac{\langle t_i^2 \rangle \langle x_i \rangle - \langle t_i \rangle \langle x_i t_i \rangle}{\langle t_i^2 \rangle - \langle t_i \rangle^2}.$$

These formulas give the *regression coefficients* used in all sciences and beyond!

7.4 Orthogonal Subspaces

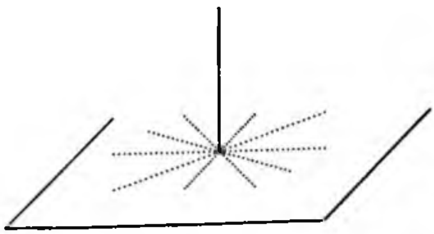
Definition. *We say that two subspaces V_1, V_2 of an inner-product space E are* orthogonal *when all vectors of V_1 are orthogonal to all vectors of V_2*

$$(\mathbf{v}_1 \mid \mathbf{v}_2) = 0 \quad \textit{for all } \mathbf{v}_1 \in V_1 \textit{ and } \mathbf{v}_2 \in V_2.$$

Taking $\mathbf{v} = \mathbf{v}_1 = \mathbf{v}_2 \in V_1 \cap V_2$, we see that $\|\mathbf{v}\|^2 = (\mathbf{v}_1 \mid \mathbf{v}_2) = 0$, hence $\mathbf{v} = 0$:

Two orthogonal subspaces have intersection $\{0\}$.

Example. In $\mathbf{R}^3$, a line and a plane (containing the origin) are orthogonal in the preceding sense precisely when all vectors in the plane are orthogonal to a generator of the line: This is the usual definition of orthogonality of a line and a plane.



But two planes of $\mathbf{R}^3$ are never orthogonal in the preceding sense since their intersection contains a line.

7.4.1 Orthogonal of a Subset

For every subset $X \subset E$, we define the *orthogonal* of X as follows

$$X^{\perp} = \{\mathbf{y} \in E : \mathbf{y} \perp \mathbf{x} \text{ for all } \mathbf{x} \in X\}$$
$$= \bigcap_{\mathbf{x} \in X} \ker(\mathbf{y} \mapsto (\mathbf{x} \mid \mathbf{y})).$$

Hence $X^{\perp}$ is a *subspace* of E.

Proposition 1. *Let E be an inner-product space and V a* finite-dimensional *subspace of E. Then $V^{\perp}$ is a supplement of V: There is an orthogonal sum decomposition $E = V \oplus V^{\perp}$ and $V^{\perp\perp} = (V^{\perp})^{\perp} = V$.*

PROOF. The first part is a reformulation of the best approximation theorem (7.3.2). Indeed, if $\mathbf{x} \in E$, we may take its best approximation $\mathbf{y}$ in V, so that $\mathbf{x} - \mathbf{y} = \mathbf{z}$ is orthogonal to V and $\mathbf{x} = \mathbf{y} + \mathbf{z} \in V + V^{\perp}$. The sum map is the inverse of $\mathbf{x} \mapsto (\mathbf{y}, \mathbf{z})$, and we have the following isomorphisms

$$\begin{array}{ccccc} E & \xrightarrow{\sim} & V \oplus V^{\perp} & \xrightarrow{\sim} & E = V + V^{\perp} \\ \mathbf{x} & \longmapsto & (\mathbf{y}, \mathbf{z}) & \longmapsto & \mathbf{x} = \mathbf{y} + \mathbf{z}. \end{array}$$

Now let us prove that the orthogonal of $V^{\perp}$ is V. Take any $\mathbf{x} \in E$ and write it $\mathbf{x} = \mathbf{y} + \mathbf{z}$ with $\mathbf{y} \in V$ and $\mathbf{z} \in V^{\perp}$. Quite generally

$$(\mathbf{x} \mid \mathbf{z}) = (\mathbf{y} + \mathbf{z} \mid \mathbf{z}) = \underbrace{(\mathbf{y} \mid \mathbf{z})}_{=0} + (\mathbf{z} \mid \mathbf{z}).$$

If now $\mathbf{x} \perp V^{\perp}$, we have $\mathbf{x} \perp \mathbf{z}$ and the preceding equalities reduce to

$$0 = (\mathbf{x} \mid \mathbf{z}) = \underbrace{(\mathbf{y} \mid \mathbf{z})}_{=0} + (\mathbf{z} \mid \mathbf{z}),$$

whence

$$\|\mathbf{z}\|^2 = (\mathbf{z} \mid \mathbf{z}) = 0, \quad \mathbf{z} = 0.$$

In this case $\mathbf{x} = \mathbf{y}$ is in V. ■

Corollary 1. *Let $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ be a finite subset of E. Then*

$$A^{\perp\perp} = (A^{\perp})^{\perp} = \mathcal{L}(A)$$

is the smallest subspace *containing A, hence is the linear span of A.*

PROOF. For any subspace V containing A, we have

$$A \subset V \quad \Longrightarrow \quad A^{\perp} \supset V^{\perp} \quad \Longrightarrow \quad A^{\perp\perp} \subset V^{\perp\perp} = V.$$ ■

In other words, the following conditions for a vector $\mathbf{b} \in E$ are equivalent:

(*i*) $\mathbf{b}$ *is a linear combination of the* $\mathbf{a}_i$

(*ii*) $\mathbf{b}$ *is orthogonal to all* $\mathbf{v} \in A^{\perp}$

(*iii*) $\mathbf{v} \perp \mathbf{a}_i$ $(1 \leqslant i \leqslant n)$ *implies* $\mathbf{v} \perp \mathbf{b}$.

Corollary 2. *Let $\mathbf{v} \neq 0$ and $V_1 = \{\mathbf{v}\}^{\perp}$. Then every $\mathbf{x}$ orthogonal to V_1 is a multiple of $\mathbf{v}$ and $E = \mathbf{R}\mathbf{v} \oplus V_1$.* ■

Corollary 3. *Let V be a finite-dimensional subspace of an inner-product space E. If $V \neq E$, then there is a unit vector $\mathbf{v} \in E$ orthogonal to V.* ∎

Comment. Notice that the finiteness assumption cannot be dropped from the preceding statements. For example, let $E = \mathcal{C}(I)$ be the space of continuous functions on a closed and bounded interval $I \subset \mathbf{R}$, with inner product defined by

$$(f \mid g) = \int_I f(x)g(x)\,dx.$$

Then, it is shown in a calculus course that the only continuous function f which is orthogonal to all polynomials is the zero function $f = 0$. (This result is not obvious: It is follows from the approximation theorem of Weierstrass.)

Proposition 2. *Any family of orthogonal subspaces of an inner-product space is independent.*

PROOF. We have to show that a finite sum $\sum_i \mathbf{v}_i$ of mutually orthogonal vectors $\mathbf{v}_i \perp \mathbf{v}_j$ $(i \neq j)$ can only vanish if all components $\mathbf{v}_i$ are zero. Assuming that $\sum_i \mathbf{v}_i = 0$, we can estimate an inner product by a typical vector $\mathbf{v}_j$ of this sum

$$\textstyle\sum_i (\mathbf{v}_i \mid \mathbf{v}_j) = \left(\sum_i \mathbf{v}_i \mid \mathbf{v}_j\right) = 0.$$

By orthogonality assumption, the inner products $(\mathbf{v}_i \mid \mathbf{v}_j) = 0$ vanish for $i \neq j$. There only remains $\|\mathbf{v}_j\|^2 = (\mathbf{v}_j \mid \mathbf{v}_j) = 0$, which proves $\mathbf{v}_j = 0$. ∎

Another PROOF. If an orthogonal sum $\sum_i \mathbf{v}_i = 0$ vanishes, the Pythagorean theorem gives

$$\sum_i \|\mathbf{v}_i\|^2 = \Big\|\sum_i \mathbf{v}_i\Big\|^2 = 0.$$

Hence $\|\mathbf{v}_i\|^2 = 0$ for all i, so that $\mathbf{v}_i = 0$ for all i. ∎

7.4.2 The Support of a Linear Map

Definition. *Let E be an inner-product space and $f : E \to F$ a linear map with a finite-dimensional kernel. Then, the* support *of f is the orthogonal of the kernel of f:*

$$\operatorname{supp} f = (\ker f)^{\perp}.$$

Under the assumptions of the definition, the support of f is a supplement to $\ker f$, hence the restriction

$$f_r = f|_{\operatorname{supp} f} : \operatorname{supp} f \xrightarrow{\sim} \operatorname{im} f$$

is an isomorphism (Sec. 5.2.2) (the index "r" is a reminder for *restriction*, as well as for *rank* of f—dimension of its image—when it is finite).

Let us determine the support of the linear map $\mathbf{R}^n \to \mathbf{R}^m$ defined by a matrix A of size $m \times n$: It is also called *support of the matrix* A. Denote by R the row space of A

$$\begin{aligned} R &= \mathcal{L}(\text{rows of } A): \quad \text{subspace of } \mathbf{R}_n, \\ {}^tR &= \mathcal{L}(\text{transposed rows of } A) \\ &= \mathcal{L}(\text{columns of } {}^tA) \\ &= \operatorname{im} {}^tA: \quad \text{subspace of } \mathbf{R}^n. \end{aligned}$$

By block multiplication we get

$$\begin{aligned} A\mathbf{x} &= \begin{pmatrix} \leftarrow & \rho_1 & \rightarrow \\ & \vdots & \\ \leftarrow & \rho_m & \rightarrow \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \\ &= \begin{pmatrix} \rho_1\mathbf{x} \\ \vdots \\ \rho_m\mathbf{x} \end{pmatrix}, \end{aligned}$$

so that

$$\begin{aligned} A\mathbf{x} &= 0, \\ \rho_i\mathbf{x} &= 0 \quad (1 \leqslant i \leqslant m), \\ \rho\mathbf{x} &= 0 \quad (\rho \in R), \\ {}^t\rho \cdot \mathbf{x} &= 0 \quad ({}^t\rho \in {}^tR) \end{aligned}$$

(with the dot product in $\mathbf{R}^n$), are all equivalent. This proves

$$\mathbf{x} \in \ker A \quad \Longleftrightarrow \quad \mathbf{x} \perp {}^tR,$$

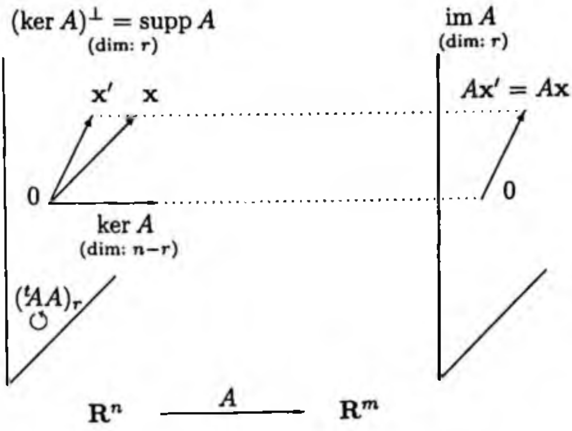
hence

$$\begin{aligned} \ker A &= ({}^tR)^\perp, \\ (\ker A)^\perp &= ({}^tR)^{\perp\perp} = {}^tR, \\ \operatorname{supp} A &= {}^tR = \operatorname{im} {}^tA. \end{aligned}$$

The equality between the row-rank and the column-rank of A is now explained by an isomorphism

$$\operatorname{supp} A = \operatorname{im} {}^tA \xrightarrow{\approx} \operatorname{im} A.$$

Let us make a picture of the linear action of A (sketching $\operatorname{im} A$ with fat lines because it is *canonical*, independent of the inner product and transposition, while $\operatorname{supp} A = \operatorname{im} {}^tA$ depends on them).



Let us summarize what we have obtained.

Proposition. *The support of A is $(\ker A)^\perp = \operatorname{im} {}^tA$, and A induces an isomorphism*

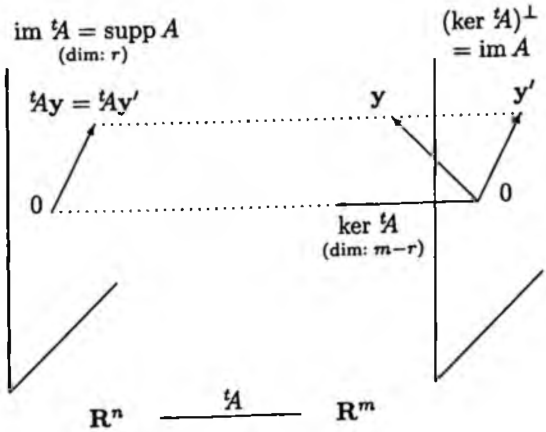
$$A_r : \operatorname{supp} A \xrightarrow{\sim} \operatorname{im} A.$$

The restriction of ${}^tAA : \mathbf{R}^n \to \mathbf{R}^m \to \mathbf{R}^n$ to the support of A is an isomorphism

$$({}^tAA)_r : \operatorname{supp} A \xrightarrow{\sim} \operatorname{supp} A.$$

■

Here is a symmetric picture for the action of ${}^tA$.



The following diagram may still help visualizing the actions of A and ${}^t\!A$:

$$
\begin{array}{c}
\mathbf{R}^n \left\{ \begin{array}{ccc} \ker A & \longrightarrow & 0 \\ \oplus & A & \\ \operatorname{im} {}^t\!A & \rightleftarrows & \operatorname{im} A \\ & {}^t\!A & \oplus \\ 0 & \longleftarrow & \ker {}^t\!A \end{array} \right\} \mathbf{R}^m .
\end{array}
$$

Corollary. *We have*

$$
\begin{aligned}
\operatorname{im} {}^t\!AA &= \operatorname{im} {}^t\!A, & \ker {}^t\!AA &= \ker A, \\
\operatorname{im} A\,{}^t\!A &= \operatorname{im} A, & \ker A\,{}^t\!A &= \ker {}^t\!A.
\end{aligned}
$$

PROOF. Since the restriction of ${}^t\!A$ gives an isomorphism $\operatorname{im} A \xrightarrow{\sim} \operatorname{im} {}^t\!A$, we have

$$\operatorname{im} {}^t\!AA = {}^t\!A(\operatorname{im} A) = \operatorname{im} {}^t\!A.$$

The inclusion $\ker {}^t\!A \subset \ker {}^t\!AA$ is obvious. Conversely, if $\mathbf{x} \in \ker {}^t\!AA$, then

$$ {}^t\!A A\mathbf{x} = 0 \quad \Longrightarrow \quad \underbrace{{}^t\mathbf{x}\,{}^t\!AA\mathbf{x}}_{\|A\mathbf{x}\|^2} = 0 \quad \Longleftrightarrow \quad A\mathbf{x} = 0, $$

and the equality $\ker {}^t\!A = \ker {}^t\!AA$ is proved. The last two equalities are obtained by replacing A by ${}^t\!A$. ■

7.4.3 Least Squares Revisited

Let us consider any linear system

$$A\,\mathbf{x} = \mathbf{b} \quad (\mathbf{x} \text{ variable in } \mathbf{R}^n,\ \mathbf{b} \text{ given in } \mathbf{R}^m), \tag{S}$$

where A is a matrix of size $m \times n$, hence defines a linear map $A : \mathbf{R}^n \to \mathbf{R}^m$. If (S) has no solution (is incompatible), it is because $\mathbf{b} \notin \operatorname{im} A$. However, the linear system

$${}^t\!AA\,\mathbf{x} = {}^t\!A\,\mathbf{b} \quad (\mathbf{x} \text{ variable in } \mathbf{R}^n,\ \mathbf{b} \text{ given in } \mathbf{R}^m) \tag{S'}$$

is always compatible since we have just seen that $\operatorname{im} {}^t\!AA = \operatorname{im} {}^t\!A$ (Sec. 7.4.2). Let us call

$$({}^t\!AA)_r : \operatorname{im} {}^t\!A \xrightarrow{\sim} \operatorname{im} {}^t\!A$$

the restriction of ${}^t\!AA$: If A has rank r, the size of $({}^t\!AA)_r$ is $r \times r$. Let us check that $({}^t\!AA)_r^{-1}({}^t\!A\,\mathbf{b})$ is a solution of (S'):

$$
\begin{aligned}
\mathbf{x} &= ({}^t\!AA)_r^{-1}({}^t\!A\,\mathbf{b}), \\
A\,\mathbf{x} &= A\,({}^t\!AA)_r^{-1}({}^t\!A\,\mathbf{b}) \quad (\text{instead of } \mathbf{b}), \\
{}^t\!AA\mathbf{x} &= {}^t\!AA({}^t\!AA)_r^{-1}({}^t\!A\,\mathbf{b}) = {}^t\!A\,\mathbf{b}.
\end{aligned}
$$

The general solution of (S') is obtained by adding to it any solution of the associated homogeneous system (HS'). Here is a diagram illustrating the situation:

$$\begin{array}{ccccc} \operatorname{supp} A & \xleftarrow{{}^t\!A} & \mathbf{R}^m & \ni \mathbf{b} & \\ ({}^t\!AA)_r^{-1} \downarrow & & \downarrow P & & \\ \mathbf{x} \in \operatorname{supp} A & \xrightarrow{A} & \operatorname{im} A \subset \mathbf{R}^m & \ni P\mathbf{b} = A\mathbf{x}. & \end{array}$$

We are going to prove that $\mathbf{x} = ({}^t\!AA)_r^{-1}({}^t\!A\,\mathbf{b})$ minimizes $\|A\mathbf{x} - \mathbf{b}\|$. In other words, $A\mathbf{x} = A\,({}^t\!AA)_r^{-1}({}^t\!A\,\mathbf{b})$ is the orthogonal projection of $\mathbf{b}$ in im ${}^t\!A$.

Proposition. *The orthogonal projector of* $\mathbf{R}^m$ *onto the subspace* $\operatorname{im} A$ *is*

$$P = A\,({}^t\!AA)_r^{-1}\,({}^t\!A),$$

and the orthogonal projector of $\mathbf{R}^n$ *onto the subspace* $\operatorname{supp} A = \operatorname{im} {}^t\!A$ *is*

$$Q = {}^t\!A\,(A\,{}^t\!A)_r^{-1}A.$$

PROOF. If $\mathbf{y} \in (\operatorname{im} A)^\perp = \ker {}^t\!A$, then $P\mathbf{y} = A\,({}^t\!AA)_r^{-1}({}^t\!A\mathbf{y}) = 0$. On the other hand, if $\mathbf{y} = A\mathbf{x} \in \operatorname{im} A$, then

$$P\mathbf{y} = A\,({}^t\!AA)_r^{-1}({}^t\!A\mathbf{y}) = A\,({}^t\!AA)_r^{-1}({}^t\!AA)\mathbf{x} = A\mathbf{x} = \mathbf{y}.$$

This proves that P is the orthogonal projector on $\operatorname{im} A$. Replacing A by ${}^t\!A$, we find the symmetric formula for Q. ■

Let us check algebraically that P is a *projector*

$$\begin{aligned} P^2 &= A\,({}^t\!AA)_r^{-1}({}^t\!A)\,A\,({}^t\!AA)_r^{-1}({}^t\!A) \\ &= A\,({}^t\!AA)_r^{-1}({}^t\!AA)_r\,({}^t\!AA)_r^{-1}({}^t\!A) \\ &= A\,({}^t\!AA)_r^{-1}({}^t\!A) = P. \end{aligned}$$

Moreover, since ${}^t\!AA$, $({}^t\!AA)_r$, and $({}^t\!AA)_r^{-1}$ are symmetric

$${}^tP = {}^{tt}\!A\,{}^t({}^t\!AA)_r^{-1}({}^t\!A) = A\,({}^t\!AA)_r^{-1}({}^t\!A) = P,$$

so that P is an *orthogonal* projector.

Let us reformulate the preceding result when A is *injective*. This will often be the case if A has size $m \times n$ where m is greater than n: (S) has more equations than variables. In this case, the rank of A is $r = n$.

Theorem. *Let A be a matrix of size $m \times n$ defining an* injective *linear map* $\mathbf{R}^n \to \mathbf{R}^m$. *Then* $\operatorname{supp} A = \mathbf{R}^n$, ${}^t\!AA : \mathbf{R}^n \to \mathbf{R}^n$ *is invertible, and for each given* $\mathbf{b} \in \mathbf{R}^m$, *there is a unique* $\mathbf{x} \in \mathbf{R}^n$ *for which* $\|A\mathbf{x} - \mathbf{b}\|$ *is minimal: It is*

$$\mathbf{x} = ({}^t\!AA)^{-1}({}^t\!A\,\mathbf{b}).$$

The orthogonal projector $P : \mathbf{R}^m \to \mathbf{R}^m$ *onto* $\operatorname{im} A$ *is*

$$P = A\,({}^t\!AA)^{-1}({}^t\!A).$$

■

7.5 Appendix: Finite Probability Spaces

The purpose of this appendix is to explain the *algebraic independence* notion occurring in probability theory, and its relation to orthogonality in inner-product spaces. It also shows the relevance of multiplying vectors component by component, an operation not considered in linear algebra.

Definition. *A* finite probability space (Ω, pr) *is a pair consisting of a finite set* Ω *and a map* $\mathrm{pr}: \Omega \to \mathbf{R}_{>0}$ *such that* $\sum_{\omega\in\Omega} \mathrm{pr}(\omega) = 1$.

By definition, Ω cannot be empty and if $\Omega = \{\omega\}$ is a singleton, then $\mathrm{pr}(\omega) = 1$. But if Ω has more than one element, the values of pr satisfy $0 < \mathrm{pr}(\omega) < 1$. In the context of finite probability spaces, the set Ω is called *sample space*, the elements of Ω are the *outcomes*, or elementary outcomes, and the subsets of Ω are the *events*. We define the *probability of an event* $A \subset \Omega$ by the formula

$$\boldsymbol{P}(A) = \sum_{\omega\in A} \mathrm{pr}(\omega) \quad (A \subset \Omega).$$

If $A = \{\omega\}$ is a singleton set, then $\boldsymbol{P}(A) = \mathrm{pr}(\omega)$: $\boldsymbol{P}$ extends the probability pr.

7.5.1 Random Variables

Let (Ω, pr) be a finite probability space.

Definition. *A* random variable *is a function* $f : \Omega \to \mathbf{R}$. *The* expectation *of* f *is the sum*

$$\boldsymbol{E}(f) = \sum_{\omega\in\Omega} \mathrm{pr}(\omega) f(\omega).$$

The set of random variables is a vector space with respect to the usual multiplication by scalars and addition of functions. If $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$, then a random variable is completely determined by its values

$$X_1 = f(\omega_1),\ X_2 = f(\omega_2), \ldots,\ X_n = f(\omega_n),$$

that we consider as the components of a vector. The vector space of random variables, also denoted by $\mathbf{R}^\Omega$, is thus identified with $\mathbf{R}^n$

$$f \longleftrightarrow X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix},$$

$$\mathcal{F}(\Omega; \mathbf{R}) \longleftrightarrow \mathbf{R}^n \quad (n : \text{cardinality of } \Omega).$$

The usual multiplication of functions, defined by $fg(\omega) = f(\omega)g(\omega)$, corresponds to the multiplication in $\mathbf{R}^n$

$$\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_1Y_1 \\ X_2Y_2 \\ \vdots \\ X_nY_n \end{pmatrix}.$$

The expectation of the random variable f is by definition the sum

$$\boldsymbol{E}(f) = \sum_{1 \leqslant i \leqslant n} \mathrm{pr}(\omega_i) X_i.$$

In the sum, we may gather all outcomes ω such that $f(\omega)$ is equal to a fixed value λ

$$\{f = \lambda\} = \{\omega \in \Omega : f(\omega) = \lambda\} = f^{-1}(\lambda),$$

and write

$$\boldsymbol{E}(f) = \sum_{\lambda \in f(\Omega)} \lambda \boldsymbol{P}\{f = \lambda\}.$$

The expectation $\boldsymbol{E} : \mathbf{R}^{\Omega} \to \mathbf{R}$ is linear, namely

$$\boldsymbol{E}(af + bg) = a\boldsymbol{E}(f) + b\boldsymbol{E}(g) \quad (a, b \in \mathbf{R};\ f, g \in \mathbf{R}^{\Omega}).$$

Since the expectation of a random variable f is a linear combination of its values, with positive coefficients

$$\boldsymbol{E}(f) \geqslant 0 \quad \text{if} \quad f \geqslant 0.$$

This implies that

$$\boldsymbol{E}(f) \leqslant \boldsymbol{E}(g) \quad \text{if} \quad f \leqslant g$$

(observe that $\boldsymbol{E}(g) - \boldsymbol{E}(f) = \boldsymbol{E}(g - f) \geqslant 0$ since $g - f \geqslant 0$). A random variable f is said to be *centered* if its expectation is 0, namely if

$$\boldsymbol{E}(f) = \sum_{\omega \in \Omega} \mathrm{pr}(\omega) f(\omega) = \sum_{\lambda \in f(\Omega)} \lambda \boldsymbol{P}\{f = \lambda\} = 0.$$

We shall consider the space of random variables $\mathbf{R}^{\Omega}$ as a Euclidean space with the inner product

$$(f \mid g) = \boldsymbol{E}(fg) = \sum_{\omega \in \Omega} \mathrm{pr}(\omega) f(\omega) g(\omega).$$

In this space, the constants make up a one-dimensional subspace generated by the constant $\boldsymbol{1}$ (the vector X having all components equal to 1). A centered random variable is simply a function orthogonal to the constant $\boldsymbol{1}$. The constant random variable taking the value c is $c\boldsymbol{1}$ (the vector having all its components equal to c). It is simply denoted by c when there is no risk of confusion with the scalar c. For example, $\boldsymbol{E}(f)$ also represents the constant random variable $\boldsymbol{E}(f)\boldsymbol{1}$. Since

$$(f - \boldsymbol{E}(f) \mid \boldsymbol{1}) = \boldsymbol{E}(f) - \boldsymbol{E}(f) = 0,$$

we see that $f - \boldsymbol{E}(f)$ is orthogonal to $\boldsymbol{1}$, hence $f - \boldsymbol{E}(f) = f_c$ is centered. There is an orthogonal, direct-sum decomposition

$$\mathbf{R}^\Omega = \mathbf{R}\boldsymbol{1} \oplus \{\boldsymbol{1}\}^\perp$$

given by

$$f = \boldsymbol{E}(f) + f_c.$$

The subspace of centered random variables has dimension $n - 1$ in $\mathbf{R}^\Omega \simeq \mathbf{R}^n$.

Definition. *Two random variables f and g are called* uncorrelated *when their centered components are orthogonal.*

Note that the conditions $f_c \perp g_c$, $f_c \perp g$, $f \perp g_c$ are equivalent.

Proposition. *Two random variables f and g are uncorrelated precisely when $\boldsymbol{E}(fg) = \boldsymbol{E}(f)\boldsymbol{E}(g)$.*

PROOF. Write $f = \boldsymbol{E}(f) + f_c$ and $g = \boldsymbol{E}(g) + g_c$ so that

$$\begin{aligned} fg &= (\boldsymbol{E}(f) + f_c)(\boldsymbol{E}(g) + g_c) \\ &= \boldsymbol{E}(f)\boldsymbol{E}(g) + \underbrace{\boldsymbol{E}(f)g_c + \boldsymbol{E}(g)f_c}_{\text{centered}} + f_c g_c, \\ \boldsymbol{E}(fg) &= \boldsymbol{E}(f)\boldsymbol{E}(g) + \boldsymbol{E}(f_c g_c). \end{aligned}$$

Obviously

$$\boldsymbol{E}(fg) = \boldsymbol{E}(f)\boldsymbol{E}(g) \quad \Longleftrightarrow \quad \boldsymbol{E}(f_c g_c) = 0 \quad \Longleftrightarrow \quad f_c \perp g_c,$$

and by definition, this happens when f and g are uncorrelated. ∎

The preceding proof shows that

$$\boldsymbol{E}(f_c^2) = \boldsymbol{E}(f^2) - \boldsymbol{E}(f)^2.$$

This is the *variance* of the random variable f. Observe that the variance is positive precisely when f is nonconstant.

Example. Consider the probability space $\Omega = \{-1, 0, +1\}$ with $\mathrm{pr}(\omega) = \frac{1}{3}$ for each outcome ω (equiprobability). Consider the two random variables f and g defined by

$$\begin{aligned} f(x) &= x, \\ g(\pm 1) &= 1 \text{ and } g(0) = -2. \end{aligned}$$

They are centered. By definition $fg = f$ is centered, so that

$$(f \mid g) = \boldsymbol{E}(fg) = \boldsymbol{E}(f) = 0 : \quad f \perp g.$$

Hence f and g are uncorrelated. But $g = 3f^2 - 2$, so that g depends on f in an obvious sense. With the vector notation

$$f \longleftrightarrow X = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \quad g \longleftrightarrow Y = \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix} = 3\begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} - 2\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

7.5.2 Algebras of Random Variables

We intend to consider the functions—in particular the polynomials—of a random variable $X \in \mathbf{R}^\Omega$. For this purpose, it is useful to introduce the notion of algebra of random variables.

Definition. *A subset $\mathcal{A} \subset \mathbf{R}^\Omega$ is a* subalgebra *when $\mathcal{A}$ is a vector subspace that contains the constant $\mathbf{1}$ and that is stable under multiplication of its elements.*

A subalgebra is characterized by the two properties

- *$\mathcal{A}$ contains the constants $c\mathbf{1}$* $\quad (c \in \mathbf{R})$,
- *f and $g \in \mathcal{A} \quad \Longrightarrow \quad f + g$ and $fg \in \mathcal{A}$,*

since they imply

- $f \in \mathcal{A},\ c \in \mathbf{R} \quad \Longrightarrow \quad cf \in \mathcal{A}$,

so that $\mathcal{A}$ is necessarily a vector subspace of $\mathbf{R}^\Omega$.

When $\mathcal{A}$ is an algebra of random variables, we denote by $\mathcal{A}_c = \mathcal{A} \cap \{\mathbf{1}\}^\perp$ the subspace consisting of the centered f in $\mathcal{A}$. Since

$$f \in \mathcal{A} \quad \Longrightarrow \quad f_c = f - \boldsymbol{E}(f)\mathbf{1} \in \mathcal{A},$$

we see that $\mathcal{A}_c$ is also the projection of $\mathcal{A}$ in $\mathbf{R}^\Omega_c$.

Here is a general construction of subalgebras of $\mathbf{R}^\Omega$. Let $\mathcal{P}$ be a partition of Ω: $\mathcal{P}$ is a family of disjoint nonempty subsets with union Ω. The random variables $f \in \mathbf{R}^\Omega$ which are constant on all elements of the partition $\mathcal{P}$ make up a subalgebra of $\mathbf{R}^\Omega$. This subalgebra will be denoted by $\mathcal{A}_\mathcal{P}$. For $A \in \mathcal{P}$, consider the corresponding characteristic function

$$\mathbf{1}_A = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Then these characteristic functions constitute a basis of the vector space $\mathcal{A}_{\mathcal{P}}$. Indeed, if a function f takes the constant value $f(A)$ on $A \in \mathcal{P}$, we have

$$f = \sum_{A \in \mathcal{P}} f(A) \mathbf{1}_A.$$

This shows that the dimension of $\mathcal{A}_{\mathcal{P}}$ is equal to the cardinality of $\mathcal{P}$.

Theorem. *Every subalgebra $\mathcal{A}$ of $\mathbf{R}^\Omega$ is of the form $\mathcal{A}_{\mathcal{P}}$ for a suitable partition $\mathcal{P}$ of the fundamental space Ω.*

PROOF. Consider the equivalence relation

$$\omega \sim \omega' \iff f(\omega) = f(\omega') \text{ for all } f \in \mathcal{A}.$$

Let $\mathcal{P}$ be the partition of Ω consisting in the equivalence classes of this relation. By definition, every $f \in \mathcal{A}_{\mathcal{P}}$ is constant on the elements of the partition $\mathcal{P}$ hence $\mathcal{A} \subset \mathcal{A}_{\mathcal{P}}$. Let us show that this inclusion is an equality. It is enough to prove that the characteristic functions $\mathbf{1}_A$ of the equivalence classes A are in $\mathcal{A}$. Now let $A \in \mathcal{P}$ be a fixed equivalence class. If $A = \Omega$, there is nothing to prove. Otherwise, for each equivalence class $B \neq A$ there is a random variable $f_B \in \mathcal{A}$ taking different values on the subsets A and B. Consider

$$\begin{aligned} g_B &= \frac{f_B - f_B(B)}{f_B(A) - f_B(B)} \\ &= a\, f_B - b\,\mathbf{1} \in \mathcal{A}. \end{aligned}$$

This function takes the value 1 on A and 0 on B. The product $\prod_{B \neq A} g_B$ corresponding to choices of $f_B \in \mathcal{A}$ is in the subalgebra $\mathcal{A}$, takes the value 1 on A and 0 on all other equivalence classes:

$$\mathbf{1}_A = \prod_{B \neq A} g_B \in \mathcal{A}.$$

This ends the proof. ∎

Example. Let $f : \Omega \to \mathbf{R}$ be a random variable. The smallest subalgebra containing f (and 1) contains all powers f^i ($i \geqslant 0$) of f and their linear combinations: These are the polynomials in f. Since the sum and product of polynomials are again polynomials, this smallest subalgebra containing f consists precisely of the polynomials in f. Observe that distinct polynomials may lead to the same function (for example, if f only takes the values 0 and 1, then $f = f^2$). We shall denote by $\mathbf{R}[f]$ this subalgebra. I claim that the partition corresponding to $\mathbf{R}[f]$ consists in the events

$$A_\lambda = \{f = \lambda\} = f^{-1}(\lambda) \quad (\lambda \in \Lambda = f(\Omega) \subset \mathbf{R}).$$

Obviously, each polynomial in f is constant on all these events. Hence the equivalence relation introduced in the proof of the theorem is

$$\omega \sim \omega' \iff f(\omega) = f(\omega')$$

for the given generator f. In particular, the dimension of $\mathbf{R}[f]$ is equal to the cardinality of $f(\Omega) \subset \mathbf{R}$.

For any $F : \mathbf{R} \to \mathbf{R}$ we may consider the composite random variable

$$F(f) = F \circ f : \Omega \xrightarrow{f} \mathbf{R} \xrightarrow{F} \mathbf{R}.$$

For example, if $F(x) = x^i$ $(i \geqslant 0)$, then $F(f) = f^i$. More generally, if F is a polynomial, $F(f)$ is the corresponding polynomial in f. The composite only depends on the restriction of F to $f(\Omega)$ and for example, f^i can be defined for negative integers as soon as 0 is not a value taken by f.

Corollary 1. *Let f be a random variable and $F : \mathbf{R} \to \mathbf{R}$ be any function. Then there exists a polynomial p such that $F(f) = p(f)$. Any random variable g which is constant on the subsets $\{f = \lambda\}$ $(\lambda \in f(\Omega))$ is a polynomial in f.* ■

Corollary 2. *Any subalgebra $\mathcal{A}$ of $\mathbf{R}^\Omega$ is generated by one random variable, hence is of the form $\mathcal{A} = \mathbf{R}[f]$ for a suitable $f \in \mathbf{R}^\Omega$.*

PROOF. By the theorem $\mathcal{A} = \mathcal{A}_\mathcal{P}$ for some partition $\mathcal{P}$. Now choose any f constant on the elements of $\mathcal{P}$, but with distinct values on these subsets. Hence $f \in \mathcal{A}_\mathcal{P}$ and

$$\mathbf{R}[f] \subset \mathcal{A}_\mathcal{P}.$$

This inclusion is an equality since

$$\dim \mathbf{R}[f] = \operatorname{card} \mathcal{P} = \dim \mathcal{A}_\mathcal{P}.$$ ■

7.5.3 Independence of Random Variables

Here is the probabilistic definition of independence for random variables.

Definition. *Two random variables f and g are called* independent *when*

$$\mathbf{R}[f]_c \perp \mathbf{R}[g]_c.$$

In other words, two random variables f and g are called independent when all polynomials (or functions) in f are uncorrelated with all polynomials (or functions) in g. This is an *algebraic condition*, much stronger than *linear independence* of f_c and g_c.

Let us give an equivalent formulation for the independence of two random variables. Since the characteristic functions $\mathbf{1}_A$ of the events $A = \{f = \lambda\}$ (resp. $\mathbf{1}_B$ of the events $B = \{g = \mu\}$) form a basis of $\mathbf{R}[f]$ (resp. $\mathbf{R}[g]$), the functions $\mathbf{1}_{A,c}$ form a set of generators of $\mathbf{R}[f]_c$ (resp. the functions $\mathbf{1}_{B,c}$ form a set of generators of $\mathbf{R}[g]_c$) and we see that f and g are independent precisely when the functions

$$\mathbf{1}_{A,c} = \mathbf{1}_A - \boldsymbol{P}(A)\,\mathbf{1} \quad (A = \{X = \lambda\})$$

are orthogonal to the functions

$$\mathbf{1}_{B,c} = \mathbf{1}_B - \boldsymbol{P}(B)\,\mathbf{1} \quad (B = \{Y = \mu\}).$$

The inner product of these functions is easily computed

$$(\mathbf{1}_{A,c} \mid \mathbf{1}_{B,c}) = (\mathbf{1}_A \mid \mathbf{1}_B) - \mathbf{P}(A)(\mathbf{1} \mid \mathbf{1}_B) \underbrace{- \mathbf{P}(B)(\mathbf{1}_A \mid \mathbf{1}) + \mathbf{P}(A)\mathbf{P}(B)}_{=0}$$
$$= (\mathbf{1}_A \mid \mathbf{1}_B) - \mathbf{P}(A)\mathbf{P}(B).$$

Since

$$(\mathbf{1}_A \mid \mathbf{1}_B) = \mathbf{E}(\mathbf{1}_A \mathbf{1}_B) = \mathbf{E}(\mathbf{1}_{A\cap B}) = \mathbf{P}(A \cap B),$$

we have

$$\mathbf{1}_{A,c} \perp \mathbf{1}_{B,c} \iff \mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B).$$

Definition. *Two events A and B are called* independent *when*

$$\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B),$$

or equivalently when their characteristic functions $\mathbf{1}_A$, $\mathbf{1}_B$ are uncorrelated

$$\mathbf{1}_{A,c} \perp \mathbf{1}_{B,c}.$$

The previous comments prove the following characterization.

Theorem. *Two random variables f and g are independent precisely when*

$\{f = \lambda\}$ and $\{g = \mu\}$ are independent for every pair $(\lambda, \mu) \in f(\Omega) \times g(\Omega)$. ■

Comment. When χ is a characteristic function, it takes the values 0 and 1 only, so that $\chi^2 = \chi$ and $\mathbf{R}[\chi]$ has dimension $\leqslant 2$: $\dim \mathbf{R}[\chi]_c \leqslant 1$. Characteristic functions are uncorrelated precisely when they are independent. But this is the only case where the property of being uncorrelated (for two random variables) is equivalent to independence.

7.6 Exercises

1. Compute the angle of the diagonal of a cube with one adjacent edge.

2. (*a*) What is the angle between the two lines

$$d_1 : x = y,\ z = 0,$$
$$d_2 : z = y,\ x = 0,$$

in the usual space $\mathbf{R}^3$? (Make a picture!)
(*b*) Same question for the two lines

$$d_1 : x = y = z,$$
$$d_2 : z = y + x - 1 = 0.$$

3. Give the matrix of the orthogonal projector on $x_1 = x_2 = x_3 = x_4$ in $\mathbf{R}^4$.

4. Consider the plane of equation $x - 2y - z = 0$ in the usual space $\mathbf{R}^3$. Find the matrices (in the canonical basis of $\mathbf{R}^3$) of the vertical projector on this plane, resp. of the orthogonal projector on the same plane. Do these matrices commute?

5. Give an orthonormal basis of the subspace of $\mathbf{R}^4$ defined by the homogeneous system

$$\begin{cases} x_1 + x_2 - 3x_3 - x_4 = 0 \\ x_1 - 3x_2 + x_3 - x_4 = 0 \\ x_1 - x_2 - x_3 - x_4 = 0. \end{cases}$$

Find the matrix of the orthogonal projector onto this subspace (in the canonical basis of $\mathbf{R}^4$).

6. For a fixed vector $\vec{a} \in \mathbf{R}^3$, find all eigenvectors and eigenvalues of the linear map

$$\vec{x} \longmapsto \vec{a} \wedge \vec{x} : \quad \mathbf{R}^3 \longrightarrow \mathbf{R}^3.$$

7. If $\vec{a}$, $\vec{b} \in \mathbf{R}^3$ are two given vectors, the equation $\vec{a} \wedge \vec{x} = \vec{b}$ represents a linear system (S) for the components x_i of $\vec{x}$. Write down this linear system and—by row operations—find the conditions under which it has a solution. Formulate the compatibility conditions in an equivalent geometric form. What is the geometrical meaning of the associated homogeneous system (HS) $\vec{a} \wedge \vec{x} = 0$?

8. Fix $\mathbf{a} \in \mathbf{R}^n$, $\mathbf{b} \in \mathbf{R}^m$ and define a linear map $T_{\mathbf{a},\mathbf{b}} : \mathbf{R}^n \longrightarrow \mathbf{R}^m$ by $T_{\mathbf{a},\mathbf{b}}(\mathbf{x}) = (\mathbf{a} \cdot \mathbf{x})\,\mathbf{b}$. (*a*) Determine the matrix of $T_{\mathbf{a},\mathbf{b}}$ with respect to the canonical bases of $\mathbf{R}^n$ and $\mathbf{R}^m$.
(*b*) When $m = n$, find all eigenvectors and eigenvalues of the operator $T_{\mathbf{a},\mathbf{b}}$.

9. Let $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ be three elements of an inner-product space, with

$$\|\mathbf{a} + \mathbf{b} + \mathbf{c}\|^2 = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 + \|\mathbf{c}\|^2.$$

Does it imply that these elements are orthogonal?

10. Let $a_1, \ldots, a_n \in \mathbf{R}$ be positive. Prove

$$\left(\textstyle\sum_i a_i\right)\left(\textstyle\sum_i 1/a_i\right) \geqslant n^2.$$

When does the equality hold?

11. Prove

$$\int_0^{\pi} \sqrt{t \sin t}\, dt < \pi.$$

12. Let $\mathbf{u}_1, \ldots, \mathbf{u}_n$, $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be elements of an inner-product space E. Consider the $n \times n$ matrix

$$P = \begin{pmatrix} (\mathbf{u}_1 \mid \mathbf{v}_1) & \cdots & (\mathbf{u}_1 \mid \mathbf{v}_n) \\ \vdots & & \vdots \\ (\mathbf{u}_n \mid \mathbf{v}_1) & \cdots & (\mathbf{u}_n \mid \mathbf{v}_n) \end{pmatrix} \qquad \text{(Gram matrix)}$$

having the inner products $p_{ij} = (\mathbf{u}_i \mid \mathbf{v}_j)$ as entries. (*a*) Assume that in an orthonormal system $(\mathbf{e}_k)$, we have

$$\mathbf{u}_i = \sum_k a_{ik}\mathbf{e}_k, \quad \mathbf{v}_j = \sum_\ell b_{j\ell}\mathbf{e}_\ell.$$

Show $P = A\,{}^tB$ where $A = (a_{ik})$ and $B = (b_{j\ell})$. Conclude that if the elements $\mathbf{u}_i$ and $\mathbf{v}_j$ belong to a subspace of dimension d, the rank of P is at most d.
(*b*) When $\mathbf{u}_i = \mathbf{v}_i$ $(1 \leqslant i \leqslant n)$, show that $\sum_{ij} p_{ij} \geqslant 0$.

13. A 3×3 regular system can be written in the form

$$(S): \quad \vec{\mathbf{a}} \cdot \vec{\mathbf{x}} = \alpha, \quad \vec{\mathbf{b}} \cdot \vec{\mathbf{x}} = \beta, \quad \vec{\mathbf{c}} \cdot \vec{\mathbf{x}} = \gamma,$$

where the three vectors $\vec{\mathbf{a}}$, $\vec{\mathbf{b}}$, and $\vec{\mathbf{c}}$ are independent. Solve the simpler system

$$(S_1): \quad \begin{cases} \vec{\mathbf{a}} \cdot \vec{\mathbf{x}} = 1 \\ \vec{\mathbf{b}} \cdot \vec{\mathbf{x}} = 0 \\ \vec{\mathbf{c}} \cdot \vec{\mathbf{x}} = 0. \end{cases}$$

Show that the solution of (S) is obtained by a superposition:

$$\vec{\mathbf{x}} = \tfrac{1}{D}(\alpha\, \vec{\mathbf{b}} \wedge \vec{\mathbf{c}} + \beta\, \vec{\mathbf{c}} \wedge \vec{\mathbf{a}} + \gamma\, \vec{\mathbf{a}} \wedge \vec{\mathbf{b}}).$$

14. In the Euclidean space $\mathbf{R}^5$, let $V = \mathcal{L}(\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ where

$$\mathbf{v}_1 = {}^t(1\;1\;1\;2\;1), \quad \mathbf{v}_2 = {}^t(1\;0\;0\;1\;-2), \quad \mathbf{v}_3 = {}^t(2\;1\;-1\;0\;2).$$

What are $\dim V$, $\dim V^\perp$? Find an orthogonal basis of $V^\perp$.

15. Let $\vec{\mathbf{u}}_i$ $(1 \leqslant i \leqslant 4)$ be four vectors in $\mathbf{R}^3$. Assume that $\|\vec{\mathbf{u}}_i - \vec{\mathbf{u}}_j\| \geqslant \delta$ for all pairs $i \neq j$. Prove that $(\vec{\mathbf{u}}_i \mid \vec{\mathbf{u}}_j) \leqslant 1 - \delta^2/2$ for the same pairs. Conclude that $\delta^2 \leqslant 8/3$ (compute $\|\sum_i \vec{\mathbf{u}}_i\|^2$), and the equality $\delta^2 = 8/3$ holds precisely when the extremities of the $\vec{\mathbf{u}}_i$ are the vertices of a regular tetrahedron (in this case $\sum_i \vec{\mathbf{u}}_i = 0$).

16. Let $(\mathbf{e}_i)_{1 \leqslant i \leqslant 4}$ be the canonical basis of $\mathbf{R}^4$, and define

$$\mathbf{v}_j = 4\mathbf{e}_j - \textstyle\sum_i \mathbf{e}_i \in V \quad (1 \leqslant j \leqslant 4),$$

where V is the 3-dimensional subspace $\{\mathbf{w}\}^\perp$, $\mathbf{w} = \sum_i \mathbf{e}_i = {}^t(1,1,1,1) \in E$ (the subspace V is also the kernel of the linear form $\varphi : (x_i) \mapsto \sum x_i$). Compute the norms of these vectors, as well as their mutual angles. Conclude that the extremities of the vectors $\mathbf{v}_j$ are the vertices of a regular tetrahedron in the Euclidean space V (isometric to $\mathbf{R}^3$). What are the lengths of the edges of this regular tetrahedron? What is the radius of the sphere inscribed in the regular tetrahedron having unit edges?

17. In an inner-product space, prove the equivalence

$$\mathbf{u} \perp \mathbf{v} \quad \Longleftrightarrow \quad \|\mathbf{u}\| \leqslant \|\mathbf{u} + a\mathbf{v}\| \text{ for all scalars } a.$$

18. In an inner-product space, two vectors $\mathbf{x}$ and $\mathbf{y}$ satisfy

$$\|\mathbf{x}\| = 10, \quad \|\mathbf{x}+\mathbf{y}\| = 11, \quad \|\mathbf{x}-\mathbf{y}\| = 9.$$

What is then $\|\mathbf{y}\|$? Make a picture.

19. Let E be the vector space of matrices of size $m \times n$. Show that the formula

$$(A \mid B) = \operatorname{tr}({}^t\!A\, B)$$

defines an inner product in E. Show that for this inner product, the symmetric matrices are orthogonal to the skew-symmetric ones.

20. Give an orthonormal basis of the vector space generated by the functions

$$1,\ \cos t,\ \sin t,\ \cos^2 t,$$

with respect to the inner product

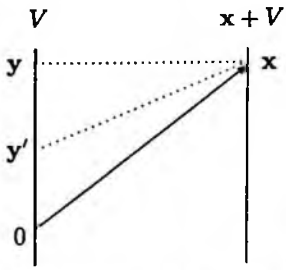
$$(f \mid g) = \frac{1}{2\pi}\int_0^{2\pi} f(t)g(t)\,dt.$$

21. Give an orthogonal basis $(P_i)_{0\leqslant i\leqslant 3}$ of the vector space consisting of the polynomial functions $f = a + bx + cx^2 + dx^3$ with inner product

$$(f \mid g) = \int_{-1}^{+1} f(x)g(x)\,dx.$$

Choose P_i of degree i and use the normalization $P_i(1) = 1$ (Legendre polynomials).

22. Let E be a Euclidean space and let V be a subspace of E. Show that if we define the distance from $\mathbf{x} \in E$ to V as $d(\mathbf{x}, V) = \inf_{\mathbf{y}\in V}\|\mathbf{x}-\mathbf{y}\| = \min_{\mathbf{y}\in V}\|\mathbf{x}-\mathbf{y}\|$, then $d(\mathbf{x}+\mathbf{v}, V) = d(\mathbf{x}, V)$ for all $\mathbf{v} \in V$.



$$\min_{\mathbf{y}'\in V}\|\mathbf{x}-\mathbf{y}'\| = \|\mathbf{x}-\mathbf{y}\|$$

Moreover, $\|\mathbf{x}+V\| = d(\mathbf{x}, V)$ defines a norm on E/V and this space is isomorphic and isometric to $V^\perp$ (see exercise 10 of Chapter 5 for the definition of the quotient space E/V).

23. Let T be an operator in an inner-product space E, and $\mathbf{v}$ a unit vector in E. Show

$$\mathbf{v} \text{ eigenvector of } T \iff (\mathbf{v} \mid T\mathbf{v})^2 = (T\mathbf{v} \mid T\mathbf{v}).$$

Notes

The last theorem in Sec. 7.2.3 leads to a necessary and sufficient condition for a *metric space* to be isometric to a subset of an inner-product space (Schoenberg's theorem). A proof can be found in

P. de la Harpe and A. Valette: La propriété (T) de Kazhdan (...), Société Mathématique de France, Astérisque 175 (1989), p.63.

Chapter 8

Symmetric Operators

To compute the powers of the matrix $\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$ (which appeared in the context of the Fibonacci numbers, see Sec. 6.4.2), we diagonalized it. The fact that this special matrix could be diagonalized came as a happy surprise. We shall now explain why this possibility was not shear luck! The main goal of this chapter is the proof of the following important result:

Any real symmetric matrix can be diagonalized (over the real field).

For example, the symmetric matrices $\begin{pmatrix} a & b \\ b & a \end{pmatrix}$ can all be diagonalized, and the same method as for the Fibonacci matrix may be used to compute their powers. In fact, symmetry of a real, square matrix is a necessary and sufficient condition for the existence of an orthogonal basis in which it acquires a diagonal form.

8.1 Definition and First Properties

A square matrix $A = (a_{ij})$ is called *symmetric* when it is equal to its transpose

$$a_{ij} = a_{ji} \quad (1 \leqslant i, j \leqslant n).$$

Let us find a translation of this equality, which is independent of a choice of basis. As we know, the columns of A are the components of the images of the vectors of the canonical basis of $\mathbf{R}^n$:

$$\begin{aligned} A\mathbf{e}_j &= j\text{th column of } A \\ a_{ij} &= \mathbf{e}_i \cdot A\mathbf{e}_j. \end{aligned}$$

The equalities $a_{ij} = a_{ji}$ are equivalent to

$$\mathbf{e}_i \cdot A\mathbf{e}_j = \mathbf{e}_j \cdot A\mathbf{e}_i = A\mathbf{e}_i \cdot \mathbf{e}_j.$$

By linearity of the inner product in its first variable, for a linear combination $\mathbf{x} = \Sigma_i x_i \mathbf{e}_i$, we deduce

$$\mathbf{x} \cdot A\mathbf{e}_j = \Sigma_i x_i \mathbf{e}_i \cdot A\mathbf{e}_j = \Sigma_i x_i A\mathbf{e}_i \cdot \mathbf{e}_j = A\Sigma_i x_i \mathbf{e}_i \cdot \mathbf{e}_j = A\mathbf{x} \cdot \mathbf{e}_j.$$

Proceeding similarly for the second variable with $\mathbf{y} = \Sigma_j y_j \mathbf{e}_j$, we find

$$\begin{aligned} \mathbf{x} \cdot \Sigma_j y_j A\mathbf{e}_j &= A\mathbf{x} \cdot \Sigma_j y_j \mathbf{e}_j, \\ \mathbf{x} \cdot A\mathbf{y} &= A\mathbf{x} \cdot \mathbf{y} \quad (\mathbf{x}, \mathbf{y} \in \mathbf{R}^n). \end{aligned}$$

8.1.1 Intrinsic Characterization of Symmetry

Our preliminary comments motivate the following definition.

Definition. *Let T be an operator in an inner-product space E. We say that T is* symmetric *when*

$$(T\mathbf{x} \mid \mathbf{y}) = (\mathbf{x} \mid T\mathbf{y}) \quad (\mathbf{x}, \mathbf{y} \in E).$$

Example. Let $\mathbf{e}$ be a unit vector. Then the operator

$$P : \mathbf{x} \longmapsto (\mathbf{x} \mid \mathbf{e})\,\mathbf{e},$$

is symmetric. Indeed,

$$(P\mathbf{x} \mid \mathbf{y}) = \big((\mathbf{x} \mid \mathbf{e})\,\mathbf{e} \mid \mathbf{y}\big) = (\mathbf{x} \mid \mathbf{e})(\mathbf{e} \mid \mathbf{y})$$

is symmetric in $\mathbf{x}$ and $\mathbf{y}$, hence equal to $(P\mathbf{y} \mid \mathbf{x}) = (\mathbf{x} \mid P\mathbf{y})$. More generally, any finite sum of rank one operators of the preceding type

$$\mathbf{x} \longmapsto P(\mathbf{x}) = \sum_{i \in J} (\mathbf{x} \mid \mathbf{e}_i)\mathbf{e}_i,$$

is a symmetric operator.

Proposition. *Let T be a symmetric operator in an inner-product space E. Then, in any orthonormal basis $(\mathbf{e}_i)$ of E, the matrix of T is symmetric.*

PROOF. Indeed, if $\mathrm{Mat}_{(e)}(T) = (a_{ij})$, we have

$$\begin{aligned} a_{ij} &= (\mathbf{e}_i \mid T\mathbf{e}_j) && (i\text{th component of } T\mathbf{e}_j \text{ given by inner product}) \\ &= (T\mathbf{e}_i \mid \mathbf{e}_j) && (T \text{ is symmetric}) \\ &= (\mathbf{e}_j \mid T\mathbf{e}_i) && (\text{symmetry of the inner product}) \\ &= a_{ji} && (j\text{th component of } T\mathbf{e}_i \text{ given by inner product}), \end{aligned}$$

hence the result. ∎

If T is a symmetric operator in E, it follows immediately from the definition that

if V is a subspace of E such that $T(V) \subset V$,
then the restriction of T to V is symmetric.

Using the proposition, we conclude that if we start with a symmetric matrix A of size $n \times n$, and a subspace $V \subset \mathbf{R}^n$ such that $A(V) \subset V$, then the restriction of A to V is given by a symmetric matrix in any orthonormal basis of V.

8.1.2 General Properties of Symmetric Operators

From the definition of symmetry, we may easily derive the basic properties of these operators.

Proposition 1. *Let T be a symmetric operator in an inner-product space E. Then the kernel and the image of T are orthogonal subspaces of E. If E is a Euclidean space, there is a direct-sum decomposition*

$$E = \ker T \oplus \operatorname{im} T.$$

PROOF. If $\mathbf{x} \in \ker T$, we have

$$(\mathbf{x} \mid T\mathbf{y}) \overset{!}{=} (T\mathbf{x} \mid \mathbf{y}) = 0 \quad (\mathbf{y} \in E),$$

hence $\mathbf{x} \perp \operatorname{im} T$. This proves $\ker T \perp \operatorname{im} T$. In particular $\ker T \cap \operatorname{im} T = \{0\}$, and these subspaces are independent. When the dimension of E is finite, the rank-nullity theorem

$$\dim E = \dim \ker T + \dim \operatorname{im} T,$$

shows that the union of bases of $\ker T$ and $\operatorname{im} T$ is a basis of E. Hence every element $\mathbf{x} \in E$ is a sum of an element $\mathbf{y} \in \ker T$ and an element $\mathbf{z} \in \operatorname{im} T$, and this decomposition is unique since these subspaces are independent. ■

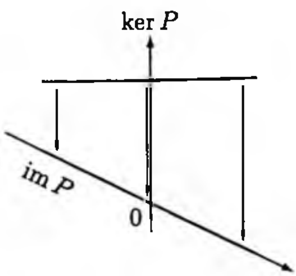
Corollary. *Orthogonal projectors are characterized by the relations*

$$P^2 = P \quad \textit{and} \quad P \textit{ symmetric.}$$

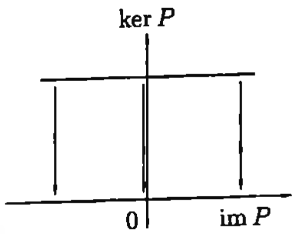
PROOF. If P is a projector, then any $\mathbf{x} \in E$ can be written $x = P\mathbf{x} + (\mathbf{x} - P\mathbf{x})$ where $\mathbf{x} - P\mathbf{x} \in \ker P$. When P is an orthogonal projector, we deduce

$$(\mathbf{x} \mid P\mathbf{y}) = (P\mathbf{x} \mid P\mathbf{y}) \quad (\mathbf{y} \in E),$$

hence P is symmetric. The converse follows from the proposition. ■



Nonsymmetric projector



Symmetric projector

Proposition 2. *If λ and μ are two distinct eigenvalues of a symmetric operator T, then the corresponding eigenspaces V_λ and V_μ are orthogonal.*

PROOF. If $T\mathbf{x} = \lambda\mathbf{x}$ and $T\mathbf{y} = \mu\mathbf{y}$, we deduce

$$\begin{aligned}(T\mathbf{x} \mid \mathbf{y}) &\overset{1}{=} (\mathbf{x} \mid T\mathbf{y}),\\ (\lambda\mathbf{x} \mid \mathbf{y}) &= (\mathbf{x} \mid \mu\mathbf{y}),\\ \lambda(\mathbf{x} \mid \mathbf{y}) &= \mu(\mathbf{x} \mid \mathbf{y}),\end{aligned}$$

hence $(\lambda - \mu)(\mathbf{x} \mid \mathbf{y}) = 0$. Since by assumption $\lambda - \mu \neq 0$, we infer $(\mathbf{x} \mid \mathbf{y}) = 0$, namely $\mathbf{x} \perp \mathbf{y}$. ∎

8.2 Diagonalization

8.2.1 Statement of the Result

We now turn to the statement of the *Spectral Theorem.*

Theorem. *Let T be an operator in a Euclidean space E. Then T is symmetric precisely when E has an orthonormal basis consisting of eigenvectors of T.*

One implication is easily seen. If E has an orthonormal basis consisting of eigenvectors of T, the operator T is represented by a diagonal matrix in this basis. This matrix is symmetric and hence so is T.

The converse, namely

> *if T is a symmetric operator in a Euclidean space E, then E has an orthonormal basis consisting of eigenvectors of T,*

is both deep and important. Several methods for constructing eigenvectors of symmetric operators are available. None of them is completely elementary. We shall give a proof in two steps (Secs. 8.2.2, 8.2.3), based on two fundamental analytical results:

- ➢ *A continuous function $f : S \to \mathbf{R}$ defined on a* closed *and* bounded *subset S of a Euclidean space, attains a maximum*
- ➢ *A differentiable function $f : I \to \mathbf{R}$ which attains a maximum at an interior point of an interval I, has a zero derivative at that point.*

We ask the reader to accept these statements. In a calculus course, the first stated result concerning continuous numerical functions defined on a closed and bounded interval of $\mathbf{R}$ is proved. The same property holds for *closed* and *bounded* subsets of a Euclidean space E. We shall use it for the unit sphere $\|\mathbf{x}\| = 1$ in E. In Sec. 12.3.3, we shall give a second independent proof of the existence of eigenvectors for symmetric operators: It is based on the fundamental theorem of algebra.

Let T be a symmetric operator in a Euclidean space E. From the spectral theorem, it follows that the set $\sigma \subset \mathbf{R}$ of eigenvalues of T is not empty. For each eigenvalue $\lambda \in \sigma$, the corresponding eigenspace is $V_\lambda = \ker(T - \lambda I) \neq \{0\}$, and these eigenspaces are mutually orthogonal. More precisely there is an orthogonal direct-sum decomposition

$$E = \bigoplus_{\lambda \in \sigma} V_\lambda.$$

The kernel of T is $\ker T = V_0$. I claim that the image of T is

$$\bigoplus_{0 \neq \lambda \in \sigma} V_\lambda.$$

Indeed, if $\mathbf{w} = \sum_{0 \neq \lambda \in \sigma} \mathbf{v}_\lambda$ ($\mathbf{v}_\lambda \in V_\lambda$), then $\mathbf{w} = T\mathbf{v}$ where

$$\mathbf{v} = \sum_{0 \neq \lambda \in \sigma} \tfrac{1}{\lambda} \mathbf{v}_\lambda.$$

We recover the orthogonality of $\ker T$ and $\operatorname{im} T$ from the orthogonality of the eigenspaces.

8.2.2 Existence of Eigenvectors

Theorem. *Any symmetric operator T in a Euclidean space E of dimension $n \geqslant 1$ has at least one eigenvector: Any unit vector $\mathbf{v}$ which maximizes the expression $(T\mathbf{x} \mid \mathbf{x})$ among unit vectors $\mathbf{x} \in E$ is an eigenvector.*

PROOF. Let us consider the numerical function $F(\mathbf{x}) = (T\mathbf{x} \mid \mathbf{x})$ on the set of unit vectors $\mathbf{x} \in E$. Take any unit $\mathbf{v}$ for which $F(\mathbf{v})$ is *maximal*. Put

$$V = \mathbf{R}\mathbf{v}, \quad W = \{\mathbf{w} \in E : \mathbf{w} \perp \mathbf{v}\} = V^\perp.$$

As the best approximation theorem shows, we have a direct-sum decomposition

$$E = V \oplus W = \mathbf{R}\mathbf{v} \oplus W,$$

and correspondingly

$$T\mathbf{v} = \lambda\mathbf{v} + \mathbf{z} \quad (\mathbf{z} \in W).$$

If $\mathbf{z} \neq 0$, we write $\mathbf{z} = \mu\mathbf{w}$ with a unit vector $\mathbf{w} \in W$ ($\mu = \|\mathbf{z}\|$). If $\mathbf{z} = 0$, we take $\mu = 0$ and choose a unit vector $\mathbf{w}$ arbitrarily in W. In all cases, we thus have

$$T\mathbf{v} = \lambda\mathbf{v} + \mu\mathbf{w}$$

with two unit orthogonal vectors $\mathbf{v}$ and $\mathbf{w}$. We are going to show that $\mu = 0$, hence $\mathbf{v}$ is an eigenvector of T:

$$T\mathbf{v} = \lambda\mathbf{v} \in \mathbf{R}\mathbf{v}.$$

The components of $T\mathbf{v}$ in the orthonormal basis $\mathbf{v}$, $\mathbf{w}$ of the two-dimensional subspace $\mathcal{L}(\mathbf{v}, \mathbf{w})$ that they generate, are given by an inner product: For example

$$(T\mathbf{v} \mid \mathbf{w}) = (\lambda\mathbf{v} + \mu\mathbf{w} \mid \mathbf{w}) = \lambda\underbrace{(\mathbf{v} \mid \mathbf{w})}_{=0} + \mu\underbrace{(\mathbf{w} \mid \mathbf{w})}_{=1} = \mu.$$

Let us consider the parametric curve

$$t \longmapsto \mathbf{x}(t) = (\cos t)\,\mathbf{v} + (\sin t)\,\mathbf{w},$$

in the subspace $\mathcal{L}(\mathbf{v},\mathbf{w})$, starting at the point $\mathbf{x}(0) = \mathbf{v}$. By the Pythagorean theorem

$$\|\mathbf{x}(t)\|^2 = \cos^2 t + \sin^2 t = 1,$$

this curve lies on the unit circle of the 2-dimensional subspace $\mathcal{L}(\mathbf{v},\mathbf{w})$, hence on the unit sphere of E. (The velocity of this parameterization is

$$\begin{aligned} \mathbf{x}'(t) &= -\sin t\ \mathbf{v} + \cos t\ \mathbf{w}, \\ \|\mathbf{x}'(t)\|^2 &= \sin^2 t + \cos^2 t = 1, \end{aligned}$$

so that it is a constant speed parameterization with initial velocity $\mathbf{x}'(0) = \mathbf{w}$.) The real-valued function (of a real variable)

$$f(t) = (\mathbf{x}(t) \mid T\mathbf{x}(t)) = F(\mathbf{x}(t))$$

has a maximum for $t = 0$ so that its derivative must vanish at this point. We have

$$\begin{aligned} \mathbf{x}(t) &= \cos t\,\mathbf{v} + \sin t\,\mathbf{w}, \\ T\mathbf{x}(t) &= \cos t\,T\mathbf{v} + \sin t\,T\mathbf{w}, \\ f(t) &= (\mathbf{v} \mid T\mathbf{v})\cos^2 t + (\mathbf{w} \mid T\mathbf{w})\sin^2 t \\ &\quad + \big((\mathbf{v} \mid T\mathbf{w}) + (\mathbf{w} \mid T\mathbf{v})\big)\cos t \sin t. \end{aligned}$$

The derivatives of both $\cos^2 t$ and $\sin^2 t$ vanish at $t = 0$ while the derivative of $\sin t \cos t$ is 1 at this point. Hence

$$f'(0) = (\mathbf{v} \mid T\mathbf{w}) + (\mathbf{w} \mid T\mathbf{v}) \overset{!}{=} 2(\mathbf{v} \mid T\mathbf{w})$$

(because T is symmetric). The vanishing of this derivative proves that $T\mathbf{w} \perp \mathbf{v}$. Hence $\mu = 0$, $T\mathbf{v} = \lambda\mathbf{v}$: $\mathbf{v}$ is an eigenvector of T. ∎

Comment. As above, with $F(\mathbf{x}) = (\mathbf{x} \mid T\mathbf{x})$, if a nonzero $\mathbf{x} \in E$ is not an eigenvector of T, then F is not maximal on the unit vector $\mathbf{v}_1 = \mathbf{x}/\|\mathbf{x}\|$. There exists a unit vector $\mathbf{v}_2$ for which $F(\mathbf{v}_2) > F(\mathbf{v}_1)$. If $\mathbf{v}_2$ is still not an eigenvector of T, we may continue, defining a sequence of unit $\mathbf{v}_n$ with increasing $F(\mathbf{v}_n)$. If we manage to get a convergent sequence $(\mathbf{v}_n)_{n\geqslant 1}$ with $F(\mathbf{v}_n)$ converging to a maximum, then $\lim \mathbf{v}_n$ will be an eigenvector of T. This method of construction of eigenvectors is based on an optimization of the *Rayleigh quotients*

$$\frac{(T\mathbf{x} \mid \mathbf{x})}{(\mathbf{x} \mid \mathbf{x})} = \frac{(T\mathbf{x} \mid \mathbf{x})}{\|\mathbf{x}\|^2} \qquad (\mathbf{x} \neq 0).$$

The construction of the greatest eigenvalue of T as a conditional maximum

$$(\mathbf{x} \mid T\mathbf{x}) \ \textit{maximal under the condition}\ (\mathbf{x} \mid \mathbf{x}) = \|\mathbf{x}\|^2 = 1$$

suggests the use of the *Lagrange parameter method.* Let us indeed introduce the Lagrange function

$$L = (\mathbf{x} \mid T\mathbf{x}) - \lambda(\mathbf{x} \mid \mathbf{x}).$$

We have to find the unconditional extremes (or critical points) of this function. These occur when its gradient vanishes. Using coordinates, with T given by a matrix A, we easily find

$$\operatorname{grad}(\mathbf{x} \mid A\mathbf{x}) = A\mathbf{x} + {}^t\!A\mathbf{x}.$$

In our case, T is symmetric so that $\operatorname{grad}(\mathbf{x} \mid T\mathbf{x}) = 2T\mathbf{x}$. In particular for $T = \mathrm{id}$, $\operatorname{grad}(\mathbf{x} \mid \mathbf{x}) = 2\mathbf{x}$. Hence we see that the condition for an unconditional extremum is

$$T\mathbf{x} - \lambda\mathbf{x} = \tfrac{1}{2}\operatorname{grad} L = 0.$$

The extremes indeed occur when $T\mathbf{x} = \lambda\mathbf{x}$, namely when $\mathbf{x}$ is a unit eigenvector.

8.2.3 Inductive Construction

Theorem. *For any symmetric operator T in a Euclidean space E, there is an orthonormal basis $(\mathbf{e}_i)$ of E consisting of eigenvectors of T, hence in which the matrix of T is diagonal.*

PROOF. We shall prove this theorem by induction on the dimension of E. Any operator in a one-dimensional space is a multiplication by a scalar, hence the case $\dim E = 1$ is trivially verified. Fix an integer $n \geqslant 2$ and assume that the theorem has been established in all dimensions less than n. Consider the case of a symmetric operator T in a space E of dimension n. Choose a unit eigenvector $\mathbf{v}$ of T: This is possible thanks to Theorem 1. Consider the subspace

$$W = \{\mathbf{v}\}^{\perp} = \{\mathbf{w} \in E : \mathbf{w} \perp \mathbf{v}\}$$

of E (7.4.1). Since $\mathbf{v} \notin W$, we have $m = \dim W < n$. On the other hand, W is an *invariant subspace* of T, namely $T(W) \subset W$: If $\mathbf{w} \perp \mathbf{v}$,

$$(T\mathbf{w} \mid \mathbf{v}) \overset{!}{=} (\mathbf{w} \mid T\mathbf{v}) = (\mathbf{w} \mid \lambda\mathbf{v}) = \lambda(\mathbf{w} \mid \mathbf{v}) = 0$$

shows that $T\mathbf{w} \perp \mathbf{v}$. The restriction of T to this subspace is a symmetric operator

$$T_W : W \longrightarrow W,$$

since the definition of symmetry only refers to the inner product of E. By induction assumption, there is an orthonormal basis $(\mathbf{e}_i)_{1\leqslant i\leqslant m}$ of W consisting in eigenvectors of T_W, hence also of T. But

$$E = \mathbf{R}\mathbf{v} \oplus W,$$

and simply adding $\mathbf{e}_0 = \mathbf{v}$ to the orthonormal system $(\mathbf{e}_i)_{1\leqslant i\leqslant m}$, we get an orthonormal basis of V consisting of eigenvectors of T. Finally, we may observe

$$n = \dim E = m + 1, \quad m = n - 1. \qquad \blacksquare$$

The induction of the preceding proof is based on the fact that if we know that the symmetric operator T has a nontrivial invariant subspace, say $T(V) \subset V$, then its orthogonal $W = V^{\perp}$ is also invariant: $T(W) \subset W$. This property is satisfied by other classes of operators (Sec. 12.3.4).

8.3 Applications

8.3.1 Quadratic Forms

Let $A = (a_{ij}) \in M_n(\mathbf{R})$ ba a square matrix. We can associate to this matrix a *quadratic form* in n variables $x_1, x_2, \ldots, x_n$, namely an expression

$$Q_A(\mathbf{x}) = Q_A(x_1, \ldots, x_n) = \sum_{1 \leqslant i,j \leqslant n} a_{ij} x_i x_j.$$

This correspondence defines a map

$$M_n(\mathbf{R}) \longrightarrow \left\{ \begin{array}{c} \text{quadratic forms} \\ \text{in } n \text{ variables} \\ x_1, \ldots, x_n \end{array} \right\}.$$

The two sets under consideration are *vector spaces* and the map in question is *linear*. It is not injective: Its kernel consists of the skew-symmetric matrices ${}^t\!A = -A$,

$$a_{ji} = -a_{ij} \quad (1 \leqslant i, j \leqslant n).$$

But we know that

$$M_n(\mathbf{R}) = \left\{ \begin{array}{c} \text{symmetric} \\ \text{matrices} \end{array} \right\} \bigoplus \left\{ \begin{array}{c} \text{skew-symmetric} \\ \text{matrices} \end{array} \right\}.$$

The decomposition is explicitly given in Sec. 6.2.4:

$$A = \underbrace{\tfrac{1}{2}(A + {}^t\!A)}_{\text{symmetric}} + \underbrace{\tfrac{1}{2}(A - {}^t\!A)}_{\text{skew-symmetric}}.$$

If the quadratic form $Q(x_1, \ldots, x_n)$ is given, say

$$Q(x_1, \ldots, x_n) = \sum_{1 \leqslant i \leqslant j \leqslant n} a_{ij} x_i x_j,$$

we choose the symmetric matrix

$$\begin{pmatrix} a_{11} & \frac{1}{2}a_{12} & \cdots \\ \frac{1}{2}a_{12} & a_{22} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix},$$

in order to have

$$Q(x_1, \ldots, x_n) = (\mathbf{x} \mid A\mathbf{x}).$$

Let $(\mathbf{v}_i)_{1\leqslant i\leqslant n}$ be an orthonormal basis consisting in eigenvectors of A. If

$$\mathbf{x} = \sum x_i \mathbf{e}_i = \sum \xi_i \mathbf{v}_i,$$

we have

$$\begin{aligned} Q(x_1, \dots, x_n) &= (\mathbf{x} \mid A\mathbf{x}) \\ &= \left(\sum \xi_j \mathbf{v}_j \mid \sum \xi_i \lambda_i \mathbf{v}_i\right) \\ &= \sum_{1\leqslant i\leqslant n} \lambda_i \xi_i^2. \end{aligned}$$

This is a representation of the quadratic form Q as a linear combination of squares. Let us order the eigenvalues of A in decreasing order

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_p > 0 > \mu_1 \geqslant \cdots \geqslant \mu_q.$$

The case $A = -I_n$ shows that all eigenvalues may be negative, in which case $p = 0$ by convention. In general, the multiplicity of the zero eigenvalue is $n - (p + q)$. We can put

$$\lambda_i = 1/a_i^2, \quad \mu_j = -1/b_j^2.$$

In a suitable orthonormal basis, the quadratic form can now be written more simply as an algebraic sum of squares

$$Q(x_1, \dots, x_n) = \sum_{1\leqslant i\leqslant p} \left(\frac{\xi_i}{a_i}\right)^2 - \sum_{1\leqslant j\leqslant q} \left(\frac{\eta_j}{b_j}\right)^2.$$

8.3.2 Classification of Quadrics

The discussion of the preceding subsection allows us to give a classification of quadrics of $\mathbf{R}^n$ having a center of symmetry. Such a quadric is a hypersurface defined by an equation

$$Q(x_1, \dots, x_n) = \sum_{1\leqslant i\leqslant j\leqslant n} a_{ij} x_i x_j = c,$$

or more simply by

$$Q(\mathbf{x}) = (\mathbf{x} \mid A\mathbf{x}) = c \quad (c \neq 0).$$

The unit sphere

$$x_1^2 + x_2^2 + \cdots + x_n^2 = 1$$

(unit circle when $n = 2$, usual unit sphere in $\mathbf{R}^3$ when $n = 3$) is a simple example of quadric. We shall not be interested in *empty* quadrics such as

$$-x_1^2 - x_2^2 = 1, \quad -x_1^2 - x_2^2 - x_3^2 = 1.$$

Since a quadratic form is a homogeneous function of degree 2

$$Q(\alpha\mathbf{x}) = \alpha^2 Q(\mathbf{x}),$$

and hence

$$Q(\mathbf{x}) = c \iff Q(\alpha\mathbf{x}) = \alpha^2 c.$$

Taking $\alpha = |c|^{-1/2}$, we see that we only have to consider the two types

$$Q(\mathbf{x}) = 1 \quad \text{or} \quad Q(\mathbf{x}) = -1.$$

The preceding discussion proves the first part of the following result.

Theorem. *Let Q be a quadratic form in $\mathbf{R}^n$. Then there is an orthonormal basis of $\mathbf{R}^n$ for which the centered quadric given by $Q(x_1, \ldots, x_n) = 1$ becomes*

$$\sum_{1 \leqslant i \leqslant p} (\xi_i/a_i)^2 - \sum_{i>p} (\xi_i/a_i)^2 = 1$$

in the new coordinates. If $p \geqslant 1$ and Q corresponds to the symmetric matrix $A \in M_n(\mathbf{R})$, namely $Q(\mathbf{x}) = (\mathbf{x} \mid A\mathbf{x})$, the greatest eigenvalue λ_1 of A is positive and the minimal half-axis of the quadric is

$$a_1 = 1/\sqrt{\lambda_1}.$$

The assumption that the quadric is nonempty is equivalent to saying that

$$(\mathbf{x} \mid A\mathbf{x}) = c^2 > 0 \quad \text{for some vector } \mathbf{x},$$

and replacing $\mathbf{x}$ by $\mathbf{y} = \mathbf{x}/c$, we can write

$$\begin{aligned} \lambda_1 &= \max_{\|\mathbf{x}\|=1} (\mathbf{x} \mid A\mathbf{x}) = \max_{\mathbf{x}\neq 0} \frac{(\mathbf{x} \mid A\mathbf{x})}{\|\mathbf{x}\|^2} \quad (> 0) \\ &\overset{!}{=} \max_{(\mathbf{y}|A\mathbf{y})=1} \frac{1}{\|\mathbf{y}\|^2} = \frac{1}{\min_{(\mathbf{y}|A\mathbf{y})=1} \|\mathbf{y}\|^2}. \end{aligned}$$

The minimal half-axis of the quadric is obviously

$$a_1 = \min_{(\mathbf{y}|A\mathbf{y})=1} \|\mathbf{y}\|$$

whence the assertion. ■

As the preceding discussion has shown, the geometrical nature of a quadric is easily described if we use an orthonormal basis in which the quadratic form is an algebraic sum of squares. The typical cases are p positive squares, q negative squares, where $0 \leqslant p + q \leqslant n$.

Example. The quadric in $\mathbf{R}^3$, given by the equation

$$xy + xz + yz = 1,$$

is a hyperboloid with two sheets. This surface has no intersection point with the coordinate axes since

$$y = z = 0 \implies xy + xz + yz = 0.$$

On the other hand, it cuts the coordinate planes according to hyperbolas: For example $z = 0$ leads to the curve $xy = 1$ in the xy-plane. Let us study this surface more systematically. We first construct the symmetric matrix

$$A = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix}$$

which corresponds to the quadratic form

$$Q(x, y, z) = xy + xz + yz.$$

An orthonormal basis consisting of eigenvectors of A is easily found if we observe that the sums of the rows of A are the same, namely 1, hence a first eigenvector ${}^t(1, 1, 1)$ with eigenvalue $\lambda_1 = 1$. The orthogonal plane has equation $x+y+z = 0$ and is an eigenspace: The geometric multiplicity of the second eigenvalue is 2. This second eigenvalue is $-\frac{1}{2}$ and we have a complete list of eigenvalues

$$\lambda_1 = 1 > 0 > \lambda_2 = \lambda_3 = -\tfrac{1}{2}.$$

Two orthogonal eigenvectors in the plane are

$$\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix}.$$

(Other choices are equally suitable: No uniqueness has been claimed, and the matrix A takes a diagonal form in several orthonormal bases.) An orthonormal basis of $\mathbf{R}^3$ consisting of eigenvectors of A is

$$\tfrac{1}{\sqrt{3}}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad \tfrac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}, \quad \tfrac{1}{\sqrt{6}}\begin{pmatrix} 1 \\ -2 \\ 1 \end{pmatrix},$$

In this basis, the quadratic form is simply

$$Q(x, y, z) = \widetilde{Q}(\xi, \eta, \zeta) = \xi^2 - \tfrac{1}{2}(\eta^2 + \zeta^2)$$

where the coefficients of the squares are the eigenvalues. The surface corresponding to the equation $Q(x, y, z) = 1$ is also given by

$$\xi^2 - \tfrac{1}{2}\rho^2 = 1 \quad \text{where } \rho^2 = \eta^2 + \zeta^2.$$

We recognize a surface having a revolution axis. It is a *hyperboloid with two sheets*. Its revolution axis is generated by the first eigenvector, namely the "diagonal" of the first octant. The minimal half-axis is

$$1/\sqrt{\lambda_1} = 1.$$

Remark. In the usual 3-space, if an eigenvalue is positive while the other two are negative, we find a hyperboloid with two sheets as above. When two eigenvalues are positive and one is negative, we find a hyperboloid with only one sheet. Finally, when the three eigenvalues are positive, we find an ellipsoid. If an eigenvalue only is 0, we find a cylinder based on either an ellipse, a hyperbola (or in the degenerate case, a pair of lines, in which case the quadric degenerates into a pair of planes).

Proposition. *The unit ball for any inner product in* $\mathbf{R}^n$ *is a full ellipsoid.*

PROOF. Any inner product in $\mathbf{R}^n$ is given by $(\mathbf{x} \mid \mathbf{y})_F = \mathbf{x} \cdot G\mathbf{y}$ where G is a positive-definite symmetric matrix (Theorem 2 in Sec. 7.3.1). The unit ball for this inner product is defined by

$$\|\mathbf{x}\|_F^2 = \mathbf{x} \cdot G\mathbf{x} \leqslant 1.$$

In any orthonormal basis in which G is diagonal, the quadratic form associated to G takes the form of a sum of squares, say

$$Q(\mathbf{x}) = \mathbf{x} \cdot G\mathbf{x} = \sum_i \lambda_i \xi_i^2,$$

where the eigenvalues λ_i are all positive, repeated according to their (geometric) multiplicities. The unit ball $\|\mathbf{x}\|_F \leqslant 1$ has for boundary the quadric

$$Q(\mathbf{x}) = \sum_i \lambda_i \xi_i^2 = 1,$$

namely is the full ellipsoid having 1/2-principal axes $1/\sqrt{\lambda_i}$. ∎

8.3.3 Positive Definite Operators

The construction of the eigenvalues of a symmetric matrix S has shown that the maximal eigenvalue of S is

$$\lambda_{\max} = \max_{\|\mathbf{x}\|=1} \mathbf{x} \cdot S\mathbf{x}.$$

Replacing S by $-S$, we find symmetrically that the minimal eigenvalue of S is

$$\lambda_{\min} = \min_{\|\mathbf{x}\|=1} \mathbf{x} \cdot S\mathbf{x},$$

so that

$$\lambda_{\min} \leqslant \mathbf{x} \cdot S\mathbf{x} \leqslant \lambda_{\max} \quad (\|\mathbf{x}\| = 1),$$

and both bounds are reached for some unit vector. This proves the following result.

Proposition. *Let S be a symmetric matrix of size $n \times n$, and $\sigma \subset \mathbf{R}$ the set of its eigenvalues. Then*

$$\lambda > 0 \text{ for all } \lambda \in \sigma \quad \Longleftrightarrow \quad \mathbf{x} \cdot S\mathbf{x} > 0 \quad \text{for all } \mathbf{x} \neq 0.$$

Moreover

$$\lambda \geqslant 0 \text{ for all } \lambda \in \sigma \quad \Longleftrightarrow \quad \mathbf{x} \cdot S\mathbf{x} \geqslant 0 \quad \text{for all } \mathbf{x} \in \mathbf{R}^n. \qquad \blacksquare$$

When $\mathbf{x} \cdot S\mathbf{x} > 0$ for all $\mathbf{x} \neq 0$, S is *positive definite*: $S \gg 0$. When $\mathbf{x} \cdot S\mathbf{x} \geqslant 0$ for all $\mathbf{x} \in \mathbf{R}^n$, we say that S is *positive semi-definite* and denote it by $S \geqslant 0$.

Proposition. *Let $S \geqslant 0$ be a symmetric positive semi-definite matrix. Then there is a unique square root $T = S^{1/2} \geqslant 0$ of S: $T^2 = S$. Moreover*

$$\ker S^{1/2} = \ker S, \quad \operatorname{supp} S^{1/2} = \operatorname{supp} S, \quad S \gg 0 \Longleftrightarrow S^{1/2} \gg 0.$$

PROOF. Let T be any symmetric matrix, and choose an orthonormal basis ε in which it is diagonal. Then T^2 is a diagonal matrix in the same basis, having as diagonal entries the square of the diagonal entries of T. The requirement $T^2 = S$ thus requires to take for diagonal entries of T the square roots of the eigenvalues of $S \geqslant 0$. Moreover, $T \geqslant 0$ imposes the choice of nonnegative square roots of these eigenvalues. This shows that if the space $\mathbf{R}^n$ is an orthogonal direct sum of eigenspaces $V_\lambda = \ker(S - \lambda I)$ of S $(0 \leqslant \lambda \in \sigma)$, then T has necessarily the same eigenspaces as S (distinct eigenvalues λ corresponding to distinct $\sqrt{\lambda} \geqslant 0$). The positive semi-definite square root T of S acts by multiplication by $\sqrt{\lambda} \geqslant 0$ in V_λ. It has the same kernel V_0 as S, hence also the same support as S. $\blacksquare$

In simple terms, the square root of S is given by a diagonal matrix in any basis in which S takes a diagonal form, and has diagonal entries equal to the square roots of those of S.

Let A be any matrix of size $m \times n$, corresponding to a linear map $\mathbf{R}^n \to \mathbf{R}^m$. Then ${}^t\!AA$ is a symmetric matrix of size $n \times n$. Since

$$\mathbf{x} \cdot {}^t\!AA\mathbf{x} = A\mathbf{x} \cdot A\mathbf{x} = \|A\mathbf{x}\|^2 \geqslant 0,$$

${}^t\!AA$ is positive semi-definite. If A is injective, ${}^t\!AA$ is positive definite: ${}^t\!AA \gg 0$.

Example. Consider the following matrix

$$A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & & 0 \\ 0 & -1 & 1 & \ddots & \vdots \\ & 0 & -1 & \ddots & \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \cdots & & 0 & -1 \end{pmatrix},$$

of size $(n+1)\times n$. Row operations show that $\operatorname{rank} A = n$, hence A is injective and ${}^t\!AA$ is positive definite. Here it is

$$
{}^t\!A = \begin{pmatrix} 1 & -1 & 0 & & \cdots & 0 \\ 0 & 1 & -1 & 0 & & \vdots \\ 0 & 0 & 1 & -1 & \ddots & \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & & 1 & -1 \end{pmatrix},
$$

$$
{}^t\!AA = \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & & \vdots \\ 0 & -1 & 2 & \ddots & \\ \vdots & & \ddots & \ddots & -1 \\ 0 & \dots & & -1 & 2 \end{pmatrix} \gg 0.
$$

Definition. *For any matrix A of size $m\times n$, its* absolute value $|A|$ *is the positive semi-definite matrix of size $n\times n$ defined by* $|A| = ({}^t\!AA)^{1/2} \geqslant 0$.

As we have seen, A, $|A|$, and $|A|^2 = {}^t\!AA$ have the same support. By restriction to this support (which we denote by an index r as in Sec. 7.4.3)

$$
A_r = A|_{\operatorname{supp} A} : \operatorname{supp} A \longrightarrow \operatorname{im} A \subset \mathbf{R}^m
$$

is injective. Moreover, since

$$
\begin{aligned}
\|A\mathbf{x}\|^2 &= \mathbf{x}\cdot {}^t\!AA\mathbf{x} \\
&= \mathbf{x}\cdot |A|^2\mathbf{x} \\
&= \|\,|A|\,\mathbf{x}\,\|^2,
\end{aligned}
$$

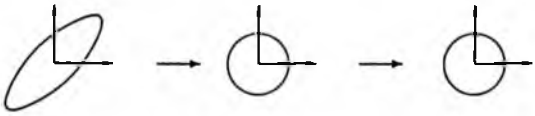
we have

$$
\begin{aligned}
\|A\mathbf{x}\| \leqslant 1 \quad &\Longleftrightarrow \quad Q(\mathbf{x}) = \mathbf{x}\cdot {}^t\!AA\mathbf{x} \leqslant 1 \\
&\Longleftrightarrow \quad \|\,|A|\,\mathbf{x}\,\| \leqslant 1.
\end{aligned}
$$

This shows that A maps a full ellipsoid of $\operatorname{supp} A$ onto the unit ball of $\operatorname{im} A$ in $\mathbf{R}^m$, while $|A|$ maps the same ellipsoid onto the unit ball of $\operatorname{supp} A$. The correspondence

$$
\begin{aligned}
J_r \;:\; \operatorname{supp} A &\longrightarrow \operatorname{im} A \\
|A|\mathbf{x} &\longmapsto A\mathbf{x},
\end{aligned}
$$

is a norm preserving linear map which will be studied more systematically in the context of isometries (Sec. 12.1.2). The restriction of A to its support has a natural factorization $A_r = J_r|A|_r$



$$\underbrace{\operatorname{supp} A \xrightarrow{|A|_r} \operatorname{supp} A \xrightarrow{J_r}}_{A_r} \operatorname{im} A \subset \mathbf{R}^m.$$

Proposition. *Let A be any matrix of size $m \times n$, corresponding to a linear map $\mathbf{R}^n \to \mathbf{R}^m$. Then with respect to the canonical Euclidean structure on these spaces, the image of the unit ball of $\mathbf{R}^n$ by A is a full ellipsoid in the subspace $\operatorname{im} A \subset \mathbf{R}^m$.*

PROOF. The inverse B of the isomorphism $\operatorname{supp} A \xrightarrow{\sim} \operatorname{im} A$ induced by A transforms a full ellipsoid $E_r \subset \operatorname{im} A$ onto the unit ball $B_r \subset \operatorname{supp} A$, hence the assertion when $\ker A = \{0\}$ (namely $\operatorname{supp} A = \mathbf{R}^n$). In general, let P be the orthogonal projector of $\mathbf{R}^n$ onto the support of A ($\ker P = \ker A$). Then

$$A : \mathbf{R}^n \xrightarrow{P} \operatorname{supp} A \xrightarrow{A_r} \operatorname{im} A \subset \mathbf{R}^m$$

transforms the unit ball B_n of $\mathbf{R}^n$ first onto the unit ball B_r of $\operatorname{supp} A$ and then onto the ellipsoid $E_r \subset \operatorname{im} A \subset \mathbf{R}^m$. (More precisely, A applies the "cylinder" $B_r \times \ker A$ onto E_r.) ■

8.4 Appendix

8.4.1 Principal Axes and Statistics

The result of scientific experiments can often be represented graphically by a cloud of points in a Euclidean space $\mathbf{R}^n$, and we may be looking for a best fit approximation in the form of a straight line. We considered this situation in the 2-plane on the occasion of the least squares method. In higher dimension, another idea is called for.

Among all lines going through the center of gravity of the set of points, we may choose the one that maximizes the variance as follows.

Introduce the vectors $\mathbf{r}_i = OP_i$ ($1 \leqslant i \leqslant N$). The center of gravity of the set of points is the extremity of

$$\mathbf{R} = \frac{1}{N} \sum_{1 \leqslant i \leqslant N} \mathbf{r}_i.$$

Replacing if necessary $\mathbf{r}_i$ by $\mathbf{r}_i{}' = \mathbf{r}_i - \mathbf{R}$, we may suppose that $\mathbf{R} = 0$:

$$\sum_{1 \leqslant i \leqslant N} \mathbf{r}_i = 0.$$

The *variance* in a given direction given by a unit vector $\mathbf{x}$ is by definition

$$f(\mathbf{x}) = \sum_{1 \leqslant i \leqslant N} (\mathbf{r}_i \mid \mathbf{x})^2.$$

We intend to find a direction in which it is maximal. But

there is a symmetric operator T for which $f(\mathbf{x}) = (\mathbf{x} \mid T\mathbf{x})$.

Consider indeed the operators

$$T_i : \mathbf{x} \longmapsto (\mathbf{x} \mid \mathbf{r}_i)\mathbf{r}_i,$$

for which

$$\operatorname{im} T_i \subset \mathbf{R}\mathbf{r}_i, \quad \operatorname{rank} T_i = \dim \operatorname{im} T_i \leqslant 1.$$

We have

$$(\mathbf{x} \mid T_i\mathbf{x}) = (\mathbf{x} \mid \mathbf{r}_i)^2.$$

These operators are symmetric

$$(T_i\mathbf{x} \mid \mathbf{y}) = ((\mathbf{x} \mid \mathbf{r}_i)\mathbf{r}_i \mid \mathbf{y}) = (\mathbf{x} \mid \mathbf{r}_i)(\mathbf{r}_i \mid \mathbf{y}).$$

Hence their sum $T = \sum T_i$ is also symmetric. The problem has been solved in the first section: The maximum occurs when the direction is given by an eigenvector of T corresponding to its largest eigenvalue.

8.4.2 Functions of a Symmetric Operator

The square root of a symmetric, positive, semi-definite operator S has already been defined. In an orthonormal basis in which S is represented by a diagonal matrix, $S^{1/2} \geqslant 0$ is also represented by a diagonal matrix.

More generally, let S be any symmetric operator in a Euclidean space. Take an orthonormal basis in which S is represented by a diagonal matrix

$$D = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix} = \operatorname{diag}(\lambda_1, \lambda_2, \dots, \lambda_n).$$

Then for any function f defined on the set $\{\lambda_i : 1 \leqslant i \leqslant n\}$ of eigenvalues, the operator $f(S)$ is defined by its matrix

$$f(D) = \begin{pmatrix} f(\lambda_1) & 0 & \dots & 0 \\ 0 & f(\lambda_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & f(\lambda_n) \end{pmatrix} = \operatorname{diag}(f(\lambda_1), f(\lambda_2), \dots, f(\lambda_n))$$

in the same basis.

For example, if $f(x) = x^n$ ($n \geqslant 1$), we obviously obtain $f(S) = S^n$. When $f = c$ is a constant, we find $f(S) = cI$. Hence for any polynomial $f = \sum_{i \leqslant n} a_i x^i$, we have $f(S) = \sum_{i \leqslant n} a_i S^i$. If S is invertible, namely does not have the eigenvalue 0, $f(x) = x^{-1}$ leads to $f(S) = S^{-1}$, inverse of S. Note however that distinct polynomials $f \neq g$ may give the same $f(S) = g(S)$. This happens precisely when $f(\lambda_i) = g(\lambda_i)$ $(1 \leqslant i \leqslant n)$. For a symmetric matrix A, there is an invertible matrix S such that $S^{-1}AS = D$ is diagonal, or $A = SDS^{-1}$. By definition, we have

$$f(A) = f(SDS^{-1}) = Sf(D)S^{-1}.$$

Proposition. *Let S and T be two operators in a Euclidean space, with T symmetric and S invertible. Then for any function f defined on the spectrum of T*

$$f(S^{-1}TS) = S^{-1}f(T)S.$$

PROOF. Let $(\mathbf{e}_i)$ be an orthonormal basis consisting of eigenvectors of T

$$T\mathbf{e}_i = \lambda_i \mathbf{e}_i \quad (1 \leqslant i \leqslant n).$$

By definition of $f(T)$ we have

$$\begin{aligned} f(T)\mathbf{e}_i &= f(\lambda_i)\mathbf{e}_i, \\ S^{-1}f(T)\mathbf{e}_i &= f(\lambda_i)S^{-1}\mathbf{e}_i, \\ S^{-1}f(T)S\varepsilon_i &= f(\lambda_i)\varepsilon_i \quad (1 \leqslant i \leqslant n). \end{aligned}$$

On the other hand,

$$(S^{-1}TS)(S^{-1}\mathbf{e}_i) = S^{-1}T\mathbf{e}_i = S^{-1}(\lambda_i \mathbf{e}_i) = \lambda_i S^{-1}\mathbf{e}_i \quad (1 \leqslant i \leqslant n)$$

proves that $S^{-1}TS$ is diagonal in the basis $(\varepsilon_i) = (S^{-1}\mathbf{e}_i)$ with the same eigenvalues λ_i. Hence by definition of $f(S^{-1}TS)$ we have

$$f(S^{-1}TS)\varepsilon_i = f(\lambda_i)\varepsilon_i \quad (1 \leqslant i \leqslant n).$$

By comparison

$$f(S^{-1}TS)\varepsilon_i = S^{-1}f(T)S\varepsilon_i \quad (1 \leqslant i \leqslant n),$$

whence $f(S^{-1}TS) = S^{-1}f(T)S$. ∎

If A is any matrix of size $m \times n$, then ${}^t\!AA$ is symmetric (of size $n \times n$) and $f({}^t\!AA)$ is well defined for any scalar function f defined on the subset $x \geqslant 0$ of $\mathbf{R}$. The square root function leads to the *absolute value* of A

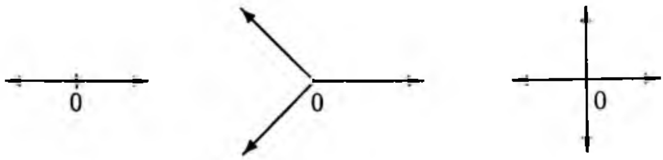
$$|A| = ({}^t\!AA)^{1/2}.$$

8.4.3 Special Configurations

Here, we shall study finite sets $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of generators of a Euclidean space such that

$$(\mathbf{v}_i \mid \mathbf{v}_j) \leqslant 0 \quad \text{for all } i \neq j.$$

Here are some examples.



The first in dimension 1 consists of two vectors with $\mathbf{v}_2 = -\mathbf{v}_1$. The other two examples, in the plane, already exhibit some characteristic features that we are going to discover in general.

Fix a set of generators $\mathbf{v}_1, \ldots, \mathbf{v}_n$ of a Euclidean space E, and consider the *Gram matrix*, namely the symmetric matrix with entries

$$g_{ij} = (\mathbf{v}_i \mid \mathbf{v}_j) \quad (1 \leqslant i, j \leqslant n).$$

The associated quadratic form is defined by

$$Q(x_1, \ldots, x_n) = \sum_{1 \leqslant i,j \leqslant n} g_{ij} x_i x_j.$$

It is positive since

$$Q(x_1, \ldots, x_n) = \left(\sum x_i \mathbf{v}_i \mid \sum x_j \mathbf{v}_j\right) = \left\|\sum x_i \mathbf{v}_i\right\|^2 \geqslant 0,$$

and moreover

$$Q(x_1, \ldots, x_n) = 0 \iff \sum_{1 \leqslant i \leqslant n} x_i \mathbf{v}_i = 0.$$

Lemma. With the assumption

$$g_{ij} = (\mathbf{v}_i \mid \mathbf{v}_j) \leqslant 0 \quad \textit{whenever } i \neq j,$$

we have

$$\sum_{1 \leqslant i \leqslant n} x_i \mathbf{v}_i = 0 \implies \sum_{1 \leqslant i \leqslant n} |x_i| \mathbf{v}_i = 0.$$

PROOF. Observe that the inequalities $x_ix_j \leqslant |x_ix_j|$ imply $g_{ij}x_ix_j \geqslant g_{ij}|x_ix_j|$ when $g_{ij} \leqslant 0$. Hence we have

$$\begin{aligned} Q(x_1,\dots,x_n) &= \sum_i g_{ii}x_i^2 + \sum_{i\neq j} g_{ij}x_ix_j \\ &\geqslant \sum_i g_{ii}x_i^2 + \sum_{i\neq j} g_{ij}\,|x_ix_j| \\ &= Q(|x_1|,\dots,|x_n|) \geqslant 0 \end{aligned}$$

by positivity of Q. As a consequence,

$$Q(x_1,\dots,x_n) = 0 \quad\Longrightarrow\quad Q(|x_1|,\dots,|x_n|) = 0,$$

and the lemma follows from the previous observation. ■

Consider now any nontrivial linear dependence relation $\sum_{1\leqslant i\leqslant n} x_i\mathbf{v}_i = 0$. Let

$$I = \{1 \leqslant i \leqslant n : x_i \neq 0\} = \{1 \leqslant i \leqslant n : |x_i| \neq 0\} \neq \varnothing.$$

Taking inner products with all $\mathbf{v}_j$, we see that this linear dependence relation implies all relations

$$\sum_{1\leqslant i\leqslant n} x_ig_{ij} = 0 \quad (1 \leqslant j \leqslant n).$$

However if we assume that $g_{ij} \leqslant 0$ for $i \neq j$, we see that when $j \notin I$

$$\sum_{i\in I} |x_i| \underbrace{g_{ij}}_{\leqslant 0} = 0 \Longrightarrow g_{ij} = 0.$$

In this case, we conclude that the family $(\mathbf{v}_i)_{i\in I}$ is orthogonal to the family $(\mathbf{v}_j)_{j\notin I}$ and E is an orthogonal sum of the two subspaces generated by these families. When $I \neq \{1,\dots,n\}$, its complement J is not empty, and we say that the configuration $\mathbf{v}_1,\dots,\mathbf{v}_n$ is *reducible*. For example, the third picture at the beginning of this section shows a reducible configuration, orthogonal sum of two configurations of the same type as the first one. It is interesting to study *irreducible* configurations, since all configurations can be built by induction from irreducible ones. Such configurations are characterized by the following property: Any partition $\{1,\dots,n\} = I \sqcup J$ such that $g_{ij} = 0$ for all $i \in I$, $j \in J$ is *trivial*, namely either $I = \varnothing$ or $J = \varnothing$.

Proposition. *Let $\mathbf{v}_1,\dots,\mathbf{v}_n$ be a set of generators of a Euclidean space E, with $g_{ij} \leqslant 0$ whenever $i \neq j$. Assume that this configuration is irreducible, namely that in any partition $\{1,\dots,n\} = I \sqcup J$ such that $g_{ij} = 0$ for all $i \in I$, $j \in J$, we have either $I = \varnothing$ or $J = \varnothing$. Then $\dim E = n$ or $n-1$. In the last case, all linear dependence relations are proportional to a basic one, having all positive coefficients.*

PROOF. If the $\mathbf{v}_i$ are independent, then $\dim E = n$. Otherwise, there is a nontrivial linear dependence relation $\sum x_i\mathbf{v}_i = 0$ and we have seen that it implies $\sum |x_i|\mathbf{v}_i = 0$. Moreover, by the irreducibility assumption,

$$J = \{1 \leqslant j \leqslant n : x_j = 0\} = \varnothing\,,$$

and all $x_i \neq 0$. Replacing all coefficients x_i by their opposite if necessary, we may assume that at least one of them is positive. In the linear dependence relation

$$\sum(|x_i| - x_i)\mathbf{v}_i = 0,$$

at least one coefficient vanishes, so that

$$I = \{1 \leqslant i \leqslant n : |x_i| \neq x_i\} \neq \{1, \ldots, n\}.$$

By the irreducibility assumption, this subset I must be empty:

$$x_i = |x_i| > 0 \quad (1 \leqslant i \leqslant n).$$

This already proves that when the $\mathbf{v}_i$ are dependent, there is a linear dependence relation $\sum x_i\mathbf{v}_i = 0$ with all $x_i > 0$. But if $\sum y_i\mathbf{v}_i = 0$ is any linear dependence relation, we may choose the scalar a in such a way that one coefficient in

$$\sum(x_i - ay_i)\mathbf{v}_i = 0$$

vanishes. As we have seen, the irreducibility assumption then implies that all coefficients vanish. Hence all linear dependence relations between the $\mathbf{v}_i$'s are proportional to the nontrivial one $\sum x_i\mathbf{v}_i = 0$. In other words, the surjective linear map

$$f : (y_i) \longmapsto \sum y_i\mathbf{v}_i, \quad \mathbf{R}^n \longrightarrow E$$

has a kernel of dimension 1. By the rank-nullity theorem,

$$\dim E = n - \dim\ker f = n - 1,$$

and the proposition is proved. ■

The four vectors

$$\vec{\mathbf{v}}_1 = \begin{pmatrix}1\\1\\1\end{pmatrix}, \quad \vec{\mathbf{v}}_2 = \begin{pmatrix}1\\-1\\-1\end{pmatrix}, \quad \vec{\mathbf{v}}_3 = \begin{pmatrix}-1\\1\\-1\end{pmatrix}, \quad \vec{\mathbf{v}}_4 = \begin{pmatrix}-1\\-1\\1\end{pmatrix},$$

give an example of an irreducible configuration in the space $\mathbf{R}^3$. (The reader should make a picture of this interesting configuration.)

8.5 Exercises

1. Find the eigenvalues and the corresponding eigenvectors of the symmetric matrix

$$\begin{pmatrix} t^2 & t & 1 \\ t & 1 & t \\ 1 & t & t^2 \end{pmatrix}.$$

2. Diagonalize the quadratic form

$$Q(x,y,z) = 7x^2 + 4xy + 6y^2 + 4yz + 5z^2$$

and determine the quadric with equation $Q(x,y,z) = 1$.

3. Determine the symmetric matrix corresponding to the quadratic form

$$Q(x,y,z) = (x-y)^2 + (y-z)^2 + (z-x)^2.$$

Is it positive definite? What is the surface $Q(x,y,z) = 1$?

4. Let A be a symmetric matrix of size $n \times n$, and $\mathbf{b} \in \mathbf{R}^n$. Consider the map $\mathbf{R}^n \to \mathbf{R}$ defined by

$$\mathbf{x} \longmapsto f(\mathbf{x}) = \tfrac{1}{2}A\mathbf{x} \cdot \mathbf{x} - \mathbf{x} \cdot \mathbf{b}.$$

Is it linear? Prove that f has an extremum at $\mathbf{x} = \mathbf{x}_0$ precisely when $A\mathbf{x}_0 = \mathbf{b}$.

5. Let T be a symmetric operator in an inner-product space E, and $\mathbf{v}$ a unit vector in E. Show

$$\mathbf{v} \text{ eigenvector of } T \iff (\mathbf{v} \mid T\mathbf{v})^2 = (\mathbf{v} \mid T^2\mathbf{v}).$$

6. Let E be the space consisting of the $n \times n$ matrices, with the inner product $(A \mid B) = \mathrm{tr}({}^tAB)$. Prove that the operator $T : A \mapsto {}^tA$ is symmetric.

7. Let

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}, \quad B = A^2 = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}.$$

(*a*) Check that the quadratic form $Q(x,y) = x^2 - xy - y^2$ is invariant under B, namely if we write it

$$Q\begin{pmatrix} x \\ y \end{pmatrix} = (x\ y)\, S \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{with } S \text{ symmetric,}$$

then

$$Q\left(B\begin{pmatrix} x \\ y \end{pmatrix}\right) = Q\begin{pmatrix} x \\ y \end{pmatrix}.$$

(*b*) Find the principal axes (and the asymptotes) of the curves $x^2 - xy - y^2 = \pm 1$.
(c) If $(f_n)_{n \geqslant 0}$ denotes the Fibonacci sequence ($f_0 = 0,\ f_1 = 1, \ldots$), prove that all integral points $P_n = \binom{f_{2n+1}}{f_{2n}}$ $(n \geqslant 0)$, are on the curve $Q(x,y) = 1$ (observe

that $BP_n = P_{n+1}$). Prove similarly that all the points $P_n = \binom{f_{2n}}{f_{2n-1}}$ $(n \geqslant 1)$, are on the curve $Q(x, y) = -1$.

8. Let T be a symmetric operator in an inner-product space E, and V a finite-dimensional subspace of E. Let $P_V : E \to V$ denote the orthogonal projection onto V (which exists by the best approximation theorem). The operator $T_V : V \to V$ *induced in* V by T is defined by $T_V = P_V T|_V$, namely by

$$T_V(\mathrm{v}) = P_V(T\mathrm{v}) \quad (\mathrm{v} \in V).$$

Show that T_V is a symmetric operator in V.

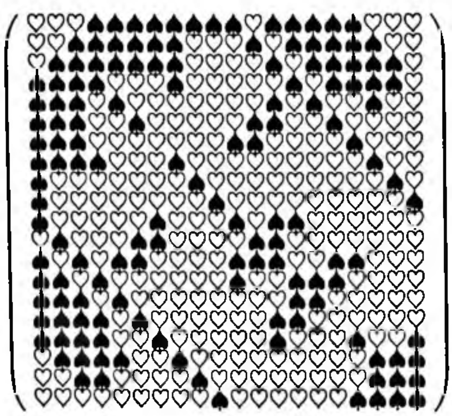
9. What are the absolute values of the matrices

$$A = \begin{pmatrix} a & -b \\ b & a \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}?$$

10. Show that the following matrices are positive definite

$$\begin{pmatrix} 4 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 4 \end{pmatrix}, \quad \begin{pmatrix} 6 & 1 & 0 & 2 \\ 1 & 6 & 1 & 0 \\ 0 & 1 & 6 & 1 \\ 2 & 0 & 1 & 6 \end{pmatrix}, \quad \begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & 0 \\ -1 & 0 & 2 \end{pmatrix}.$$

What are the eigenvalues of the last one?



Here is a symmetric matrix!

Chapter 9

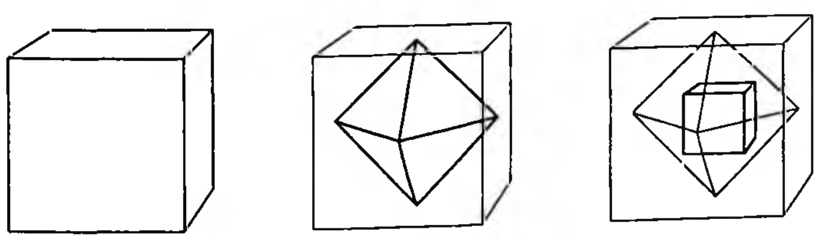
Duality

Duality is a complex notion having ramifications in analysis, geometry, algebra (to mention only mathematics). It refers to a certain symmetry of order two, similar to a mirror image. A duality has to furnish a picture in which some properties are reversed. When applied twice, the original picture should reappear.

9.1 Geometric Introduction

9.1.1 Duality for Platonic Solids

The cube has six faces and eight vertices while the octahedron has eight faces and six vertices (note that they have the same number of edges). The numbers of faces and vertices for these solids are exchanged. This exchange is an example of duality. More precisely, here is a geometrical explanation for it. Linking the center of the faces of a cube produces a regular octahedron. Similarly, linking the center of the faces of a regular octahedron furnishes a cube. Hence iterating twice this construction reproduces the initial solid, up to scale.



The preceding geometric duality construction can be made for all *Platonic solids*. Here is a list of these regular polyhedra.

	v vertices	e edges	f faces	$f-e+v$
tetrahedron	4	6	4	2
cube	8	12	6	2
octahedron	6	12	8	2
dodecahedron	20	30	12	2
icosahedron	12	30	20	2

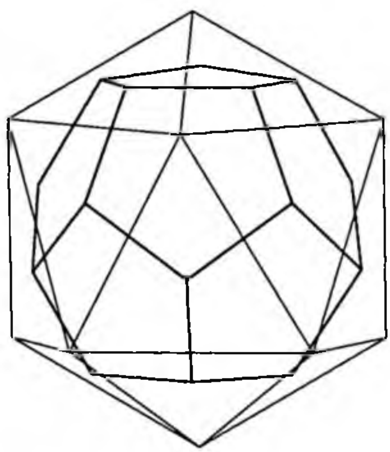
The exchange of the numbers of faces and vertices

$$f \longleftrightarrow v$$

(leaving the number e of edges invariant) produces a symmetry

$$\begin{array}{rcl} \text{tetrahedron} & \longleftrightarrow & \text{tetrahedron} \\ \text{cube} & \longleftrightarrow & \text{octahedron} \\ \text{icosahedron} & \longleftrightarrow & \text{dodecahedron.} \end{array}$$

Here is a geometrical illustration of the duality for the dodecahedron and the icosahedron.



Linear algebra provides a far-reaching generalization of the duality between the cube and the octahedron (see the comment in Sec. 9.3.3).

9.1.2 The Pappus Theorem and its Dual

Let us trace two straight lines ℓ and ℓ' in a plane. Choose three points on each of them

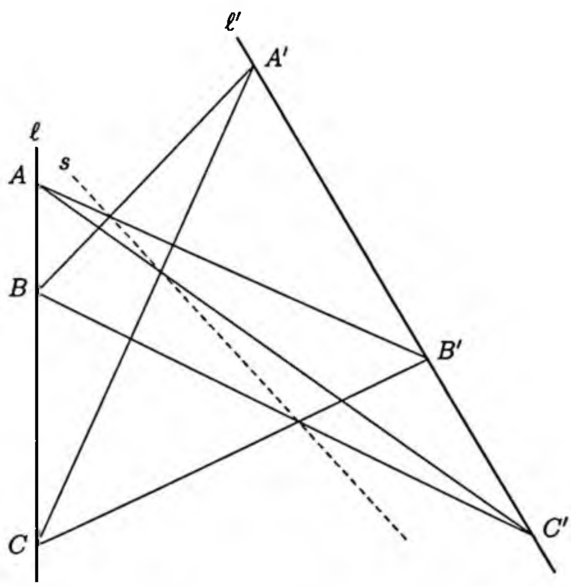
$$A,\ B,\ C\ \in\ \ell,$$
$$A',\ B',\ C'\ \in\ \ell'.$$

Let us denote generally by PQ the straight line going through two distinct points P and Q. Then the three intersection points

$$AB'\ \cap\ A'B,$$
$$AC'\ \cap\ A'C,$$
$$BC'\ \cap\ B'C$$

are aligned, say on the straight line s.

Here is an illustration of this configuration.



This is the PAPPUS theorem. To obtain its *dual*, we exchange

$$\text{point} \longleftrightarrow \text{line},$$
$$\in\ \longleftrightarrow\ \ni,$$

in the previous statement.

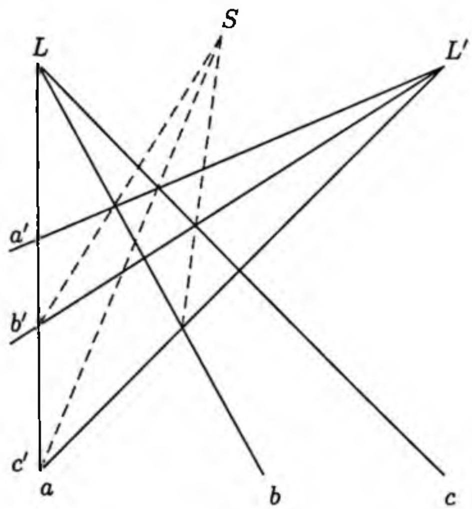
Then we obtain the following *dual theorem.* Take two points L and L' in the plane. Choose three lines going through each of these points, say

$$a,\ b,\ c \ni L; \quad a',\ b',\ c' \ni L'.$$

Then the three lines linking pairs of intersection points

$$a \cdot b' \,\&\, a' \cdot b, \quad a \cdot c' \,\&\, a' \cdot c, \quad b \cdot c' \,\&\, b' \cdot c$$

are concurrent: We may call S their common intersection point. (Observe that the notation is chosen in a good mnemonic way: Capitals for points and small letters for lines translate duality into an exchange of lower- and upper-case letters.)



In these statements, it is understood that two parallels have a common point at infinity.

We shall not prove these geometric theorems. But let us simply note that if one has been proved, the other one is automatically proved also since

the axioms of projective geometry are invariant under duality.

For example: *There is one and only one line containing a given pair of distinct points*, and dually, *two distinct lines have one and only one common point* (possibly at infinity if they are parallel). This axiom imposes that a line has a point at infinity, and this point is common to all parallel lines, but different for non parallel ones. (There is one point at infinity in each direction.)

9.2 Dual of a Vector Space

A paradigm of duality is furnished by the correspondence

$$\mathbf{R}^n\text{: column vectors} \quad \longleftrightarrow \quad \mathbf{R}_n\text{: row vectors.}$$

To understand the nature of this transposition, we view a row vector as a $1 \times n$ matrix, hence as a linear map

$$(a_1 \cdots a_n) : \mathbf{R}^n \longrightarrow \mathbf{R}.$$

This is a linear form on column vectors

$$\mathbf{x} = (x_i) \longmapsto a_1 x_1 + \cdots + a_n x_n.$$

Before studying linear forms systematically, let us illustrate their natural appearance.

A physical measure may sometimes be viewed as a linear form. Consider for example an ingot composed of unknown quantities of copper, silver, and gold: Say x kilos of copper, y kilos of silver, and z kilos of gold. A measure of its *weight* (in kilos) gives a first linear equation

$$x + y + z = A.$$

In the 3-dimensional space, this first measure indicates that the triple describing the composition of this ingot is on a certain plane. Then, we can measure the *volume* (in liters: weight = volume × density). Since the densities of the three constituents are different, we obtain a second independent relation, say

$$x/d + y/e + z/f = B.$$

Finally, the *value* (in dollars) of the ingot furnishes a third equation

$$xp + yq + zr = C$$

allowing the precise determination of the triple (x, y, z).

9.2.1 Definition and First Properties

When φ and $\psi : E \to \mathbf{R}$ are linear forms, then $a\varphi + \psi$ (a scalar) denotes the function

$$a\varphi + \psi : \mathbf{x} \longmapsto a\varphi(\mathbf{x}) + \psi(\mathbf{x}).$$

This is again a linear form, and this shows that the linear forms constitute a vector space: A vector subspace of the space of all functions $E \to \mathbf{R}$.

Definition. *The dual of a vector space E is the vector space E^* whose elements are the linear forms on E*

$$E^* = \{\varphi : E \to \mathbf{R} \text{ linear}\}.$$

If $0 \neq \mathbf{a} \in E$, there is a basis of E containing this element, and since we may define a linear map $E \to \mathbf{R}$ by prescribing arbitrarily the images of the basis vectors, there is a linear form $\varphi \in E^*$ such that $\varphi(\mathbf{a}) = 1 \neq 0$. In particular, we see that if $E \neq \{0\}$, then $E^* \neq \{0\}$.

Proposition 1. *Let $\varphi \in E^*$ be a nonzero linear form on E, and let $\mathbf{a} \in E$ be any element such that $\varphi(\mathbf{a}) \neq 0$. then there is a direct-sum decomposition*

$$\begin{aligned} E &= \mathbf{R}\mathbf{a} \oplus \ker\varphi \\ \mathbf{x} &= \ \mathbf{y} \ + \ \mathbf{z}, \end{aligned}$$

given by $\mathbf{y} = (\varphi(\mathbf{x})/\varphi(\mathbf{a}))\,\mathbf{a}$, $\mathbf{z} = \mathbf{x} - (\varphi(\mathbf{x})/\varphi(\mathbf{a}))\,\mathbf{a}$.

PROOF. It is enough to check that $\mathbf{z}$, as defined in the statement, belongs to the kernel of φ

$$\varphi(\mathbf{z}) = \varphi(\mathbf{x}) - \frac{\varphi(\mathbf{x})}{\varphi(\mathbf{a})}\varphi(\mathbf{a}) = 0.$$

The statement follows. ∎

In other words, any vector $\mathbf{a}$ not in the kernel of a linear form generates a supplement of this kernel: The kernel of a nonzero linear form has *codimension* 1. When E is finite dimensional, the rank-nullity theorem

$$\dim\ker\varphi + \dim\operatorname{im}\varphi = \dim E \quad (\varphi \in E^*),$$

also shows

$$\varphi \neq 0 \quad \Longleftrightarrow \quad \operatorname{im}\varphi = \mathbf{R} \quad \Longleftrightarrow \quad \dim\ker\varphi = \dim E - 1.$$

Corollary. *Let φ, $\psi \in E^*$. If ψ vanishes on* $\ker\varphi$*, then $\psi = \lambda\varphi$ is a multiple of φ.*

PROOF. If $\varphi = 0$, there is nothing to prove. Let us assume that $\varphi \neq 0$, and choose $\mathbf{a} \in E$ such that $\varphi(\mathbf{a}) \neq 0$. Consider the linear form $\varphi(\mathbf{a})\psi - \psi(\mathbf{a})\varphi$, which vanishes on $\mathbf{a}$ and on $\ker\varphi$ by assumption. By the proposition, this linear form vanishes identically: We get a linear dependence relation

$$\varphi(\mathbf{a})\psi - \psi(\mathbf{a})\varphi = 0,$$

from which we now deduce $\psi = \lambda\varphi$, with $\lambda = \psi(\mathbf{a})/\varphi(\mathbf{a})$. ∎

Here is a generalization.

Proposition 2. *Let $\varphi_i \in E^*$ $(1 \leqslant i \leqslant m)$ be a finite set of linear forms. If a linear form $\psi \in E^*$ vanishes on $\bigcap_{1\leqslant i\leqslant m}\ker\varphi_i$, then ψ is a linear combination of the φ_i.*

PROOF. Let us proceed by induction on m, noting that for $m = 1$, it is the statement of the above corollary. Let now $m \geqslant 2$, and consider the subspace

$$V = \bigcap_{2\leqslant i\leqslant m} \ker\varphi_i \supset V \cap \ker\varphi_1 = \bigcap_{1\leqslant i\leqslant m} \ker\varphi_i,$$

as well as the restrictions of ψ and φ_1 to V. By assumption, $\psi|_V$ vanishes on $\ker \varphi_1|_V$ hence there is a scalar λ_1 such that $\psi|_V = \lambda_1\varphi_1|_V$, namely $\psi - \lambda_1\varphi_1$ vanishes on V. By the induction assumption, $\psi - \lambda_1\varphi_1$ is a linear combination of $\varphi_2, \ldots, \varphi_m$, say

$$\psi - \lambda_1\varphi_1 = \sum_{2\leqslant i\leqslant m} \lambda_i\varphi_i,$$

hence the result. ■

9.2.2 Dual Bases

Let E be a vector space. Choose a basis $(\mathbf{e}_i)_{i\in I}$ of E. Then each $\mathbf{x} \in E$ has a unique expression $\mathbf{x} = \sum_{i\in I} x_i\mathbf{e}_i$, and for each index $j \in I$, the map

$$\varepsilon_j : \mathbf{x} \longmapsto x_j,\ E \longrightarrow \mathbf{R},$$

is linear: It is the jth *coordinate form* corresponding to the chosen basis. By definition we have

$$\varepsilon_j(\mathbf{e}_i) = \delta_{ij},$$

and these equalities characterize uniquely the coordinate forms corresponding to the basis $(\mathbf{e}_i)_{i\in I}$. We may write

$$\mathbf{x} = \sum_{i\in I} \varepsilon_i(\mathbf{x})\mathbf{e}_i.$$

For any linear form $\varphi \in E^*$, we have

$$\varphi(\mathbf{x}) = \sum_{i\in I} \varepsilon_i(\mathbf{x})\varphi(\mathbf{e}_i).$$

Let us define $a_i = \varphi(\mathbf{e}_i)$ $(i \in I)$. Hence

$$\varphi(\mathbf{x}) = \sum_{i\in I} a_i\varepsilon_i(\mathbf{x}) \quad (\mathbf{x} \in E).$$

It is tempting to write

$$\varphi = \sum_{i\in I} a_i\varepsilon_i$$

but as this may be an infinite sum, it is only a symbolic expression. However, one may be comforted by the fact that when evaluated on a specific vector $\mathbf{x} \in E$ it has only finitely many nonzero terms, since by definition of a basis, the linear combination $\mathbf{x} = \sum_{i\in I} x_i\mathbf{e}_i$ has only finitely many nonzero coefficients x_i.

Proposition. *When the dimension of E is finite, then E^* is also finite dimensional and* $\dim E^* = \dim E$. *Moreover, for each basis $(\mathbf{e}_i)_{1\leqslant i\leqslant n}$ of E, there is a unique basis $(\varepsilon_i)_{1\leqslant i\leqslant n}$ of E^* for which $\varepsilon_j(\mathbf{e}_i) = \delta_{ij}$ $(1 \leqslant i, j \leqslant n)$.*

PROOF. As in the preliminary comments, if $\varphi \in E^*$ is any linear form on E, we can write

$$\varphi = \sum_{1 \leqslant j \leqslant n} a_j \varepsilon_j \quad \text{where } a_j = \varphi(\mathbf{e}_j).$$

This proves that the family $(\varepsilon_j)_{1 \leqslant j \leqslant n}$ generates the dual E^*. There only remains to show that these coordinate forms ε_j are independent. But if $\sum_j a_j \varepsilon_j = 0$ is a linear dependence relation, we may estimate it on the basis vector $\mathbf{e}_i$

$$0 = \sum_{1 \leqslant j \leqslant n} a_j \varepsilon_j(\mathbf{e}_i) = \sum_{1 \leqslant j \leqslant n} a_j \delta_{ij} = a_i.$$

Hence all coefficients in the linear dependence relation vanish. ∎

Definition. *The basis $(\varepsilon_j)_{1 \leqslant j \leqslant n}$ of E^*, such that $\varepsilon_j(\mathbf{e}_i) = \delta_{ij}$, is the* dual basis *of $(\mathbf{e}_i)_{1 \leqslant i \leqslant n}$.*

Let us remember that the components of a linear form $\varphi \in E^*$ in the dual basis of $(\mathbf{e}_i)_{1 \leqslant i \leqslant n}$ are the values $\varphi(\mathbf{e}_i)$.

9.2.3 Bidual of a Vector Space

The construction of a dual may be iterated:

The dual of E^* is the space of linear forms on E^*.

To get a feeling for this abstract construction, let us comment on the general situation. In the first half of the twentieth century, mathematicians used to speak of a function $f(x)$ because it was understood that in this notation, the variable was x. Later on, it appeared useful to have f as a variable, often keeping $x = a$ fixed! For example, the Dirac function was precisely interpreted as an evaluation

$$f \longmapsto f(a)$$

on a particular function space. Here is a picture illustrating this interpretation

$$\left\{ \begin{matrix} \text{functions} \\ f \end{matrix} \right\} \to \boxed{\begin{matrix} \text{evaluation} \\ \varepsilon_a \end{matrix}} \to f(a).$$

We prefer to speak of a function f (not mentioning its variable), keeping $f(x)$ for the image of x under this map. This distinction is particularly important with an evaluation map ε_a having a function f as a variable: The value of this evaluation ε_a at the point f is

$$\varepsilon_a(f) = f(a).$$

In our case, we evaluate linear forms φ at a fixed vector $\mathbf{a} \in E$

$$\varphi \to \boxed{\varepsilon_{\mathbf{a}}} \to \varphi(\mathbf{a}).$$

In this way, the evaluation appears as a linear form on E^*, namely an element of the dual of this space E^*

$$\varepsilon_{\mathbf{a}} \in (E^*)^* = E^{**}.$$

Finally, letting the element $\mathbf{a}$ vary in E, we obtain a (linear!) map

$$\varepsilon : E \longrightarrow E^{**}, \quad \mathbf{a} \longmapsto \varepsilon_{\mathbf{a}}.$$

Theorem. *The linear map $\varepsilon : E \to E^{**}$ defined by $\mathbf{a} \mapsto \varepsilon_{\mathbf{a}} = (\varphi \mapsto \varphi(\mathbf{a}))$ is injective. When E is finite dimensional, it is a canonical isomorphism.*

PROOF. The meaning of $\varepsilon_{\mathbf{a}} = 0$ is $\varphi(\mathbf{a}) = \varepsilon_{\mathbf{a}}(\varphi) = 0$ for all linear forms φ. It implies $\mathbf{a} = 0$ (another application of the incomplete basis theorem!). Hence $\ker\varepsilon = \{0\}$: The linear map ε is *injective.* In the finite-dimensional case, we have seen

$$\dim E = \dim E^* = \dim E^{**}$$

and ε is automatically surjective. ∎

Quite generally, the injective linear map ε always defines a *canonical embedding* of the space E in its bidual.

9.3 Dual of a Normed Space

Recall that a normed space (Sec. 7.2.4) is a pair consisting of a real vector space E and a map $\mathbf{x} \mapsto \|\mathbf{x}\| : E \to \mathbf{R}$, satisfying:

$$\begin{array}{ll} \text{(N1)} & \|\mathbf{x}\| > 0 \text{ when } \mathbf{x} \neq 0 \\ \text{(N2)} & \|a\mathbf{x}\| = |a|\|\mathbf{x}\| \quad (a \text{ scalar}) \\ \text{(N3)} & \|\mathbf{x}+\mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|. \end{array}$$

Taking $a = 0$ in (N2), we see that $\|\mathbf{0}\| = 0$, so that $\|\mathbf{x}\| \geqslant 0$ for all $\mathbf{x} \in E$.

9.3.1 Dual Norm

Let E be a *normed space.* Let us show how one can define the norm of a linear form on E.

Example. Consider a linear form in the usual plane $\mathbf{R}^2$, say

$$\varphi : \vec{r} = \begin{pmatrix} x \\ y \end{pmatrix} \longmapsto ax + by.$$

The graph of φ is a plane in $\mathbf{R}^2 \times \mathbf{R} = \mathbf{R}^3$. If we fix a unit vector $\vec{r} \in \mathbf{R}^2$, the ratio

$$|\varphi(a\vec{r})|/\|a\vec{r}\| = |\varphi(\vec{r})|/\|\vec{r}\| = |\varphi(\vec{r})|$$

is the same for all nonzero scalars a, and represents the slope of the line generated by $\begin{pmatrix} \vec{r} \\ \varphi(\vec{r}) \end{pmatrix}$ in the graph of φ. By definition, the norm of φ is the least upper bound of all $|\varphi(\vec{r})|$ for unit vectors $\vec{r}$: We recognize the definition of the slope of a plane, here the graph of φ. Observe that the norm of φ depends on the choice of norm in $\mathbf{R}^2$. One possibility is to take the Euclidean norm

$$\|\vec{r}\| = \sqrt{x^2+y^2} \quad \text{if } \vec{r} = \begin{pmatrix} x \\ y \end{pmatrix}.$$

The case of a linear form on $\varphi : \mathbf{R}^n \to \mathbf{R}$, say $\mathbf{x} = (x_i) \mapsto \sum_i a_i x_i$, is treated similarly.

Definition. *The norm of the linear form $\varphi \in E^*$ is the least constant c for which*

$$|\varphi(\mathbf{x})| \leqslant c\|\mathbf{x}\| \quad (\mathbf{x} \in E).$$

It is denoted by $\|\varphi\|^*$ or more simply by $\|\varphi\|$.

By definition, we have an optimal general inequality

$$|\varphi(\mathbf{x})| \leqslant \|\varphi\|\,\|\mathbf{x}\| \quad (\mathbf{x} \in E).$$

Taking unit vectors $\mathbf{x} \in E$, we see that

$$|\varphi(\mathbf{x})| \leqslant \|\varphi\| \quad (\|\mathbf{x}\| = 1),$$

and since this is optimal

$$\sup_{\|\mathbf{x}\|=1} |\varphi(\mathbf{x})| = \|\varphi\|.$$

One can see similarly that

$$\|\varphi\| = \sup_{\|\mathbf{x}\|\neq 0} \frac{|\varphi(\mathbf{x})|}{\|\mathbf{x}\|}.$$

Proposition. *The map $\varphi \mapsto \|\varphi\|$ is a norm on E^*, namely, it satisfies the properties $(N1)$, $(N2)$, and $(N3)$.*

The norm just defined on E^* is called the *dual norm* with respect to the given normed space $(E, \|\cdot\|)$.

9.3.2 Dual of a Euclidean Space

Let E be a Euclidean space, namely a finite-dimensional real vector space equipped with an inner product. Each $\mathbf{v} \in E$ defines a linear form

$$\varphi_{\mathbf{v}} : \mathbf{x} \longmapsto (\mathbf{v} \mid \mathbf{x}),$$

simply since the inner product is linear in its second variable.

Theorem (Riesz). *Every linear form φ on a Euclidean space E is given by an inner product: There exists a vector $\mathbf{v} \in E$ such that*

$$\varphi(\mathbf{x}) = (\mathbf{v} \mid \mathbf{x}) \quad (\mathbf{x} \in E).$$

PROOF. As above, for $\mathbf{v} \in E$, let us define $\varphi_{\mathbf{v}}(\mathbf{x}) = (\mathbf{v} \mid \mathbf{x})$. The mapping

$$\mathbf{v} \longmapsto \varphi_{\mathbf{v}} : E \longrightarrow E^*$$

is linear. Since $\varphi_{\mathbf{v}}(\mathbf{v}) = (\mathbf{v} \mid \mathbf{v}) > 0$ if $\mathbf{v} \neq 0$, it is injective. Since E is finite dimensional, it is an isomorphism. ∎

The preceding proof has the advantage of being short, but it may be interesting to know how to find the vector $\mathbf{v}$ corresponding to a given linear form $\varphi \neq 0$. Here is a more constructive way of proceeding. There is a unit $\mathbf{n} \in E$ which is orthogonal to $\ker\varphi$ (Corollary 3 of Proposition 1 in Sec. 7.4.1). Since the inner product against this vector $\mathbf{n}$ vanishes on $\ker\varphi$ by definition, $\varphi_{\mathbf{n}}$ is proportional to φ (Corollary of Proposition 1 in Sec. 9.2.1). Explicitly, one may use the orthogonal sum decomposition

$$\begin{aligned} E &= \mathbf{R}\mathbf{n} \oplus \ker\varphi \\ \mathbf{x} &= a\mathbf{n} + \mathbf{w} \quad (a = (\mathbf{n} \mid \mathbf{x}),\ \mathbf{w} \in \ker\varphi), \end{aligned}$$

to compute φ:

$$\varphi(\mathbf{x}) = a\varphi(\mathbf{n}) = \varphi(\mathbf{n})(\mathbf{n} \mid \mathbf{x}).$$

Hence we see $\varphi = \varphi_{\mathbf{v}}$ with $\mathbf{v} = \varphi(\mathbf{n})\mathbf{n}$.

Corollary. *The normed dual of a Euclidean space E is isometric to E. More precisely, the linear mapping*

$$\begin{aligned} E &\longrightarrow E^* \\ \mathbf{v} &\longmapsto \varphi_{\mathbf{v}} = (\mathbf{v} \mid \cdot) \end{aligned}$$

is an isometry.

PROOF. We only have to check that

$$\|\varphi_{\mathbf{v}}\| = \sup_{\|\mathbf{x}\|=1} |(\mathbf{v} \mid \mathbf{x})| = \|\mathbf{v}\|.$$

But the Cauchy–Schwarz inequality gives

$$|(\mathbf{v} \mid \mathbf{x})| \leqslant \|\mathbf{v}\| \quad (\|\mathbf{x}\| = 1),$$

hence $\|\varphi_{\mathbf{v}}\| \leqslant \|\mathbf{v}\|$. Conversely if $\mathbf{v} \neq 0$ consider the unit vector $\mathbf{x} = \mathbf{v}/\|\mathbf{v}\|$ for which

$$|(\mathbf{v} \mid \mathbf{x})| = (\mathbf{v} \mid \mathbf{v})/\|\mathbf{v}\| = \|\mathbf{v}\|,$$

so that $\|\varphi_{\mathbf{v}}\| \geqslant \|\mathbf{v}\|$. ∎

If $(\mathbf{u}_i)$ is a basis of a Euclidean space E, the dual basis consists of the linear forms $\varphi_j = (\mathbf{v}_j \mid \cdot)$ such that

$$\varphi_j(\mathbf{u}_i) = (\mathbf{v}_j \mid \mathbf{u}_i) = \delta_{ij}.$$

Identifying the dual E^* to E, we also call $(\mathbf{v}_j)$ the *dual basis* of $(\mathbf{u}_i)$.

Example. Let $\vec{\mathbf{u}}_1, \vec{\mathbf{u}}_2, \vec{\mathbf{u}}_3$ be a basis of the usual Euclidean space $\mathbf{R}^3$. Then the scalar triple product $d = \vec{\mathbf{u}}_1 \cdot (\vec{\mathbf{u}}_2 \wedge \vec{\mathbf{u}}_3)$ is nonzero, and the dual basis of $(\vec{\mathbf{u}}_i)$ is

$$\begin{aligned} \vec{\mathbf{v}}_1 &= \tfrac{1}{d}\,\vec{\mathbf{u}}_2 \wedge \vec{\mathbf{u}}_3, \\ \vec{\mathbf{v}}_2 &= \tfrac{1}{d}\,\vec{\mathbf{u}}_3 \wedge \vec{\mathbf{u}}_1, \\ \vec{\mathbf{v}}_3 &= \tfrac{1}{d}\,\vec{\mathbf{u}}_1 \wedge \vec{\mathbf{u}}_2. \end{aligned}$$

A nice example of dual basis is obtained if we start with

$$\vec{\mathbf{u}}_1 = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}, \quad \vec{\mathbf{u}}_2 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad \vec{\mathbf{u}}_3 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix},$$

having vertices on the faces of the unit cube. Then $d = 2$ and the dual basis is proportional to

$$\vec{\mathbf{u}}_2 \wedge \vec{\mathbf{u}}_3 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}, \quad \vec{\mathbf{u}}_3 \wedge \vec{\mathbf{u}}_1 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}, \quad \vec{\mathbf{u}}_1 \wedge \vec{\mathbf{u}}_2 = \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}.$$

9.3.3 Dual of Important Norms in $\mathbf{R}^n$

Let us now determine the norms of linear forms in $\mathbf{R}^n$ for several norms

$$\begin{aligned} \|\mathbf{x}\|_1 &= \|(x_i)\|_1 &&= \Sigma_i |x_i|, \\ \|\mathbf{x}\|_2 &= \|(x_i)\|_2 &&= \sqrt{\Sigma_i |x_i|^2}, \\ \|\mathbf{x}\|_\infty &= \|(x_i)\|_\infty &&= \max_i |x_i|, \end{aligned}$$

on this space. Let us start by the quadratic norm: Since this one comes from the inner product

$$(\mathbf{y} \mid \mathbf{x}) = \sum_i y_i x_i, \quad \|\mathbf{x}\|_2^2 = (\mathbf{x} \mid \mathbf{x}),$$

$\mathbf{R}^n$ is a Euclidean space, isometric to its dual.

Proposition 1. The norm of a linear form $\varphi = (a_i)$ on the normed space $(\mathbf{R}^n, \|\cdot\|_2)$ is $\|(a_i)\|_2$: The dual of $(\mathbf{R}^n, \|\cdot\|_2)$ is $(\mathbf{R}_n, \|\cdot\|_2)$.

PROOF. Here, we consider $\varphi(\mathbf{x}) = \sum_i a_i x_i$ as an inner product $\varphi(\mathbf{x}) = (\mathbf{a} \mid \mathbf{x})$ and use the Cauchy–Schwarz inequality (for the quadratic norm)

$$|\varphi(\mathbf{x})| = |(\mathbf{a} \mid \mathbf{x})| \leqslant \|\mathbf{a}\|_2 \|\mathbf{x}\|_2,$$

hence $\|\varphi\| \leqslant \|\mathbf{a}\|_2$. But the choice $\mathbf{x} = \mathbf{a}$ shows that $\varphi(\mathbf{a}) = (\mathbf{a} \mid \mathbf{a}) = \|\mathbf{a}\|_2^2$, so that conversely $\|\varphi\| \geqslant \|\mathbf{a}\|_2$. This proves $\|\varphi\| = \|\mathbf{a}\|_2$. ∎

Proposition 2. *The norm of a linear form $\varphi = (a_i)$ on the normed space $(\mathbf{R}^n, \|\,.\,\|_1)$ is $\|(a_i)\|_\infty$: The dual of $(\mathbf{R}^n, \|\,.\,\|_1)$ is $(\mathbf{R}_n, \|\,.\,\|_\infty)$.*

PROOF. The obvious inequalities

$$\left|\sum_i a_i x_i\right| \leqslant \sum_i |a_i x_i| \leqslant (\max_i |a_i|) \sum_i |x_i| = \|(a_i)\|_\infty \|(x_i)\|_1$$

show that the norm of φ on $(\mathbf{R}^n, \|\,.\,\|_1)$ is smaller or equal to $\|(a_i)\|_\infty$. Choosing $\mathbf{x} = \mathbf{e}_j$ in succession, we see that

$$\|\varphi\| \geqslant |\varphi(\mathbf{e}_j)| = |a_j| \quad (1 \leqslant j \leqslant n),$$

so that the preceding estimate is optimal: $\|\varphi\| \geqslant \max_j |a_j| = \|(a_i)\|_\infty$. ■

Proposition 3. *The norm of a linear form $\varphi = (a_i)$ on the normed space $(\mathbf{R}^n, \|\,.\,\|_\infty)$ is $\|(a_i)\|_1$: The dual of $(\mathbf{R}^n, \|\,.\,\|_\infty)$ is $(\mathbf{R}_n, \|\,.\,\|_1)$.*

PROOF. The obvious inequalities (compare with the preceding proof)

$$\left|\sum_i a_i x_i\right| \leqslant \sum_i |a_i x_i| \leqslant \sum_i |a_i| \max_i |x_i| = \|(a_i)\|_1 \|(x_i)\|_\infty$$

show that the norm of φ on $(\mathbf{R}^n, \|\,.\,\|_\infty)$ is smaller or equal to $\|(a_i)\|_1$. But we cannot find a smaller bound since the vector $\mathbf{x} = (\mathrm{sgn}(a_i))$,[1] normed if $\mathbf{a} \neq 0$, also requires

$$\|\varphi\| \geqslant |\varphi(\mathbf{x})| = \sum_i a_i \,\mathrm{sgn}(a_i) = \sum_i |a_i| = \|(a_i)\|_1.$$

Hence $\|\varphi\| = \sum_i |a_i| = \|(a_i)\|_1$. ■

Comment. The unit ball for $\|\,.\,\|_1$ is an octahedron, while the unit ball for the dual norm $\|\,.\,\|_\infty$ is a cube: We recover the Platonic duality between these regular solids. The dodecahedron is the unit ball for a norm on $\mathbf{R}^3$ (which is not so easily given algebraically) whose dual norm has the icosahedron for unit ball (same remark). Notice that if B is the unit ball for a norm in a real vector space, then B is symmetric with respect to the origin in the following sense

$$\mathbf{x} \in B \iff -\mathbf{x} \in B.$$

The tetrahedron does not have this symmetry, hence is not a unit ball for a norm on $\mathbf{R}^3$. Nevertheless, a general duality theory (independent of norms) is available for convex sets. It leads to the expected duality for convex polyhedra (replace faces by vertices and conversely), and applies to the tetrahedron.

Theorem (Hahn-Banach). *In any normed space E, we have*

$$\|\mathbf{x}\| = \sup_{\varphi \neq 0} |\varphi(\mathbf{x})| / \|\varphi\|.$$ ■

This result is easy to establish in the usual space $\mathbf{R}^3$, with any norm. It is valid in any (even infinite) dimension. We shall not prove it.

Corollary. *The canonical isomorphism $E \to E^{**}$ is* isometric. ■

---

[1] The sign function sgn is defined by $\mathrm{sgn}(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0. \\ -1 & \text{if } x < 0 \end{cases}$

9.4 Transposition of Linear Maps

9.4.1 Transposition of Operators in Euclidean Spaces

Let T be an operator in a Euclidean space E.

Definition. *We say that an operator T' (in E) is a* transpose *of T when*

$$(T\mathbf{x} \mid \mathbf{y}) = (\mathbf{x} \mid T'\mathbf{y}) \quad \text{for all } \mathbf{x}, \mathbf{y} \in E.$$

Proposition. *Every operator T has exactly one transposed operator T' also denoted by ${}^tT$.*

PROOF. Let us take an orthonormal basis $(\mathbf{e}_i)$ of E, and let

$$A = (a_{ij}) = \mathrm{Mat}_{(e)}(T), \quad B = (b_{ij}) = \mathrm{Mat}_{(e)}(T').$$

Taking $\mathbf{x} = \mathbf{e}_i$ and $\mathbf{y} = \mathbf{e}_j$, we see that T' is a transpose of T precisely when $a_{ji} = b_{ij}$, hence when the matrices of T and of T' (in this basis) are transposed, whence the assertion. ∎

Since the determinant of a matrix is the same as the determinant of the transposed matrix, the preceding proof shows that an operator and its transpose have the same determinant. The *existence* of a transpose of T can also be based on the Riesz theorem (Sec. 9.3.2). To define $T'\mathbf{y}$, let us consider the linear form $\mathbf{x} \mapsto (T\mathbf{x} \mid \mathbf{y})$: It is given by an inner product against a vector, which by definition is $T'\mathbf{y}$. One can also show *uniqueness* of transposition in an intrinsic way (namely without choice of a basis) as follows: If T' and T'' are two transpose of T, we have

$$(\mathbf{x} \mid T'\mathbf{y}) = (T\mathbf{x} \mid \mathbf{y}) = (\mathbf{x} \mid T''\mathbf{y}) \quad (\mathbf{x}, \mathbf{y} \in E),$$

hence $(\mathbf{x} \mid T'\mathbf{y} - T''\mathbf{y}) = 0$ for all $\mathbf{x}, \mathbf{y} \in E$. Taking $\mathbf{x} = T'\mathbf{y} - T''\mathbf{y}$ we get $\|T'\mathbf{y} - T''\mathbf{y}\|^2 = 0$, hence $T'\mathbf{y} - T''\mathbf{y} = 0$. Since this is true for all $\mathbf{y} \in E$, it proves $T' = T''$.

Definition. *An operator T in a Euclidean space E is*

symmetric *when* ${}^tT = T$:	$(T\mathbf{x} \mid \mathbf{y}) = (\mathbf{x} \mid T\mathbf{y})$,
skew-symmetric *when* ${}^tT = -T$:	$(T\mathbf{x} \mid \mathbf{y}) = -(\mathbf{x} \mid T\mathbf{y})$,
orthogonal *when* ${}^tT = T^{-1}$:	$(T\mathbf{x} \mid \mathbf{y}) = (\mathbf{x} \mid T^{-1}\mathbf{y})$,

for all $\mathbf{x}, \mathbf{y} \in E$.

The proof of existence and uniqueness of transposition, made by use of an orthonormal basis shows that *in any orthonormal basis,*

the matrix of a symmetric operator is symmetric,
the matrix of a skew-symmetric operator is skew-symmetric,
the matrix of an orthogonal operator is orthogonal.

9.4.2 Abstract Formulation of Transposition

Just as a row vector is a matrix description of a linear map $\mathbf{R}^n \to \mathbf{R}$, we can view a column vector as an $n \times 1$ matrix representing a linear map

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} : \mathbf{R} \longrightarrow \mathbf{R}^n.$$

This is the linear map characterized by $1 \mapsto \mathbf{x}$, and is a first instance of a transpose of a linear map. More generally, let us examine the case of an $m \times n$ matrix A. Up to now, we identified A to the linear map

$$\begin{aligned} A : \mathbf{R}^n &\longrightarrow \mathbf{R}^m \\ \mathbf{x} &\longmapsto A\mathbf{x}, \end{aligned}$$

that it defines by left multiplication. But the matrix A also defines a natural linear map

$$\begin{aligned} A' : \mathbf{R}_m &\longrightarrow \mathbf{R}_n \\ (a_1, \ldots, a_m) &\longmapsto (a_1, \ldots, a_m)A, \end{aligned}$$

by right multiplication. *The matrix of this linear map with respect to the canonical bases of* $\mathbf{R}_m$ *and* $\mathbf{R}_n$ *is the transpose of* A. This is easily seen: The entries in the first column of the matrix of A' are the components of the image of the first basis vector $(1, 0, \ldots, 0)$: These entries are read in

$$(1, 0, \ldots, 0)A = \rho_1 : \text{ first row of } A.$$

Similarly, the jth column of the matrix of A' is the jth row of A, hence the statement. Since row vectors correspond to linear forms, and matrix multiplication corresponds to composition of linear maps, A' corresponds to the composition

$$\varphi \longmapsto A'(\varphi) = \varphi \circ A.$$

The following definition is an intrinsic description of the preceding particular case.

Definition. *Let* $T : E \to F$ *be any linear map. The* transpose *of* T *is the linear map*

$$T^* : F^* \longrightarrow E^*, \quad T^*(\varphi) = \varphi \circ T.$$

The notation T^* is quite natural, but will soon be replaced by ${}^tT$ or T' since T^* is traditionally reserved for the adjoint of T, to be defined in Sec. 12.3.2. By definition, we have $(T^*\varphi)x = \varphi(Tx)$ $(x \in E)$. A few diagrams may help to visualize this duality

$$\begin{array}{ccc} E & \xrightarrow{T} & F \\ \mathbf{x}\uparrow & \nearrow & Tx \\ \mathbf{R} & & \end{array} \qquad \longleftrightarrow \qquad \begin{array}{ccc} E & \xrightarrow{T} & F \\ T^*(\varphi) & \searrow & \downarrow\varphi \\ & & \mathbf{R} \end{array}$$

Hence we obtain a (linear!) map

$$\begin{aligned}\mathcal{L}(E;F) &\longrightarrow \mathcal{L}(F^*;E^*)\\ T &\longmapsto T^*.\end{aligned}$$

The following diagrams

$$E \xrightarrow{T_1} E_1 \xrightarrow{T_2} E_2,$$

$$E^* \xleftarrow{T_1^*} E_1^* \xleftarrow{T_2^*} E_2^*,$$

show why duality reverses the order of composition:

$$(T_2 \circ T_1)^* = T_1^* \circ T_2^*.$$

Proposition. *Let $T : E \to F$ be a linear map and let A be the matrix of T with respect to some bases of E and F. Then the matrix B of the transposed operator $T^* : F^* \to E^*$ with respect to the dual bases is the transposed matrix $B = {}^t\!A$.*

PROOF. To simplify notation, let us only consider the case $E = F$. By definition of the matrix A of T in the basis $(\mathbf{e}_\ell)$

$$\begin{aligned}T\mathbf{e}_j &= \textstyle\sum_\ell a_{\ell j}\mathbf{e}_\ell\\ \varepsilon_i(T\mathbf{e}_j) &= \textstyle\sum_\ell a_{\ell j}\underbrace{\varepsilon_i(\mathbf{e}_\ell)}_{=\delta_{i\ell}} = a_{ij}.\end{aligned}$$

The matrix B of T^* in the dual basis (ε_k) is similarly defined by

$$\begin{aligned}T^*\varepsilon_i &= \textstyle\sum_k b_{ki}\varepsilon_k\\ T^*\varepsilon_i(\mathbf{e}_j) &= \textstyle\sum_k b_{ki}\underbrace{\varepsilon_k(\mathbf{e}_j)}_{=\delta_{kj}} = b_{ji}.\end{aligned}$$

But by definition $\varepsilon_i(T\mathbf{e}_j) = (\varepsilon_i \circ T)(\mathbf{e}_j) = (T^*\varepsilon_i)(\mathbf{e}_j)$, so that $a_{ij} = b_{ji}$. ∎

The preceding result shows that the terminology of transpose of a linear map T is well chosen, since it extends the matrix context. Transposition has other interesting duality features.

Proposition. *For a linear map $T : E \to F$ and its transpose T^* we have:*

$$\begin{aligned}\ker T^* &= \{\psi \in F^* : \psi \textit{ vanishes on } \operatorname{im} T\}\\ \operatorname{im} T^* &= \{\varphi \in E^* : \varphi \textit{ vanishes on } \ker T\}.\end{aligned}$$

PROOF. The first assertion is trivial since

$$\psi \in \ker T^* \iff \psi \circ T = 0 \iff \psi|_{\operatorname{im} T} = 0.$$

For the second assertion, notice that $T^*\psi = \psi \circ T$ vanishes on $\ker T$. Conversely, if $\varphi \in E^*$ vanishes on $\ker T$, the restriction of φ to a supplement $W \simeq \operatorname{im} T \subset F$ of $\ker \varphi$ corresponds to a linear form on $\operatorname{im} T$ which can be extended into a linear form ψ on F (by a choice of supplement of this subspace in F). Hence $\varphi = \psi \circ T = T^*(\psi) \in \operatorname{im} T^*$. ∎

Corollary. *For the transpose T^* of a linear map $T : E \to F$, we have:*

$$\begin{aligned} T \textit{ surjective} &\iff T^* \textit{ injective} \\ T \textit{ injective} &\iff T^* \textit{ surjective.} \end{aligned}$$

■

9.5 Exercises

1. Draw the solid having vertices ${}^t(0,\pm1,\pm1)$, ${}^t(\pm1,0,\pm1)$, and ${}^t(\pm1,\pm1,0)$. How many vertices and faces does it have? What is its dual?

2. The Desargues theorem is the following statement. Let ABC and $A'B'C'$ be two triangles in the plane. If the straight lines a linking A and A', b linking B and B', and c linking C and C' go through a point S, then the intersection points I_1 of the sides AB and $A'B'$, I_2 of the sides BC and $B'C'$, and I_3 of the sides CA and $C'A'$ are aligned. What is the dual statement?

3. If a hexagon has its vertices on an ellipse, then the pairs of opposite sides intersect at three aligned points (Pascal's theorem). Formulate the dual statement concerning a hexagon having sides tangent to an ellipse (Brianchon's theorem; here, one may replace the ellipse by any conic curve).

4. Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be elements of a vector space E, $\varphi_1, \ldots, \varphi_n$ be elements of the dual E^*. Consider the $n \times n$ matrix $P = (\varphi_i(\mathbf{v}_j))$ (Gram matrix). Let $(\mathbf{e}_k)$ be a basis of E and (ε_k) the dual basis of E^*. If

$$\mathbf{v}_i = \sum_k a_{ik}\mathbf{e}_k, \quad \varphi_j = \sum_\ell b_{j\ell}\varepsilon_\ell,$$

show $P = A\,{}^tB$ where $A = (a_{ik})$ and $B = (b_{j\ell})$ (see exercise 12 of Chapter 7).

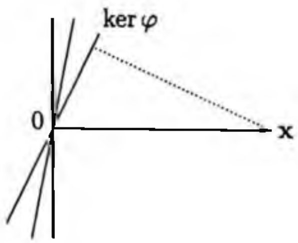
5. Let E be a Euclidean space and $0 \neq \varphi \in E^*$. Take any unit vector $\mathbf{v} \in E$, $\mathbf{v} \perp \ker\varphi$. Show that the inner product $(\varphi(\mathbf{v})\mathbf{v} \mid \cdot)$ represents the linear form φ, namely $\varphi(\mathbf{x}) = (\varphi(\mathbf{v})\mathbf{v} \mid \mathbf{x})$ (replacing $\mathbf{v}$ by $-\mathbf{v}$ does not change $\varphi(\mathbf{v})\mathbf{v}$).

6. Let E be a finite-dimensional normed space, φ a linear form on E. Recall that the norm of φ is defined by $\|\varphi\| = \sup_{\|\mathbf{x}\|\leqslant 1}|\varphi(\mathbf{x})|$. If $\varphi \neq 0$, prove

$$\min_{\varphi(\mathbf{x})=1}\|\mathbf{x}\| = 1/\|\varphi\|.$$

7. Let E be a Euclidean space. Show that if φ is a linear form on E with $\|\varphi\| = 1$, then the distance of $\mathbf{x}$ to $\ker\varphi$ is (exercise 22 of Chapter 7)

$$d(\mathbf{x}, \ker\varphi) = |\varphi(\mathbf{x})| \quad \text{and} \quad \max_{\|\varphi\|=1}|\varphi(\mathbf{x})| = \|\mathbf{x}\|.$$



$$\max_{\varphi\neq 0} d(\mathbf{x}, \ker\varphi) = \|\mathbf{x}\|$$

8. Let $\mathbf{a}$ be a nonzero element of a Euclidean space E. Consider the linear map $T : E \to \mathbf{R}$ defined by $T\mathbf{x} = (\mathbf{a} \mid \mathbf{x})$. What is the transpose of T? If $\mathbf{b}$ is any nonzero element in a vector space F, and $T_{\mathbf{a},\mathbf{b}} : E \to F$ is defined by $T_{\mathbf{a},\mathbf{b}}(\mathbf{x}) = (\mathbf{a} \mid \mathbf{x})\mathbf{b}$, what is the transpose of $T_{\mathbf{a},\mathbf{b}}$? What is the rank of this transpose?

9. Let E be any vector space. For any subset $S \subset E$ define

$$S^{\perp} = \{\varphi \in E^{*} : \varphi(S) = 0\},$$

so that $S^{\perp}$ is a vector subspace of E^{*}. Prove that for any subspace $V \subset E$, the transpose of $\pi : E \to E/V$ (see exercise 10 of Chapter 5 for the definition of E/V) gives an isomorphism

$$(E/V)^{*} \xrightarrow{\sim} V^{\perp} \subset E^{*}.$$

Notes

PAPPUS OF ALEXANDRIA is the last geometer of the Alexandrian school. He writes his *Collection* (8 books) around 320 A.D.: The quoted theorem in Sec. 9.1.2 is Proposition 139 in his Book VII.

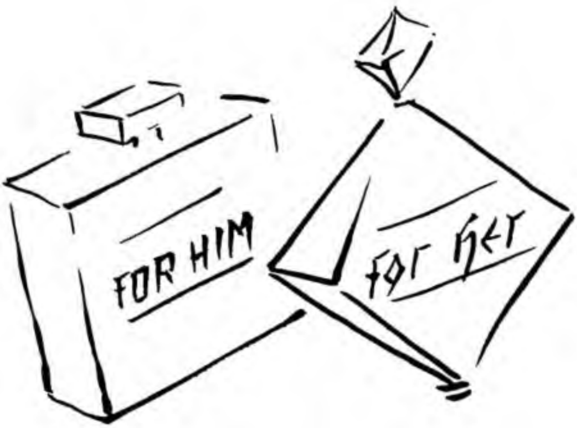
Here is an interesting visual experiment. Start with two identical hexagons

and add three segments in each to obtain dual pictures

Observe how the brain forgets the original profile in favor of the 3-dimensional representation.

Keywords for Web Search

Johnson solids
Archimedean solids and duality: Catalan solids

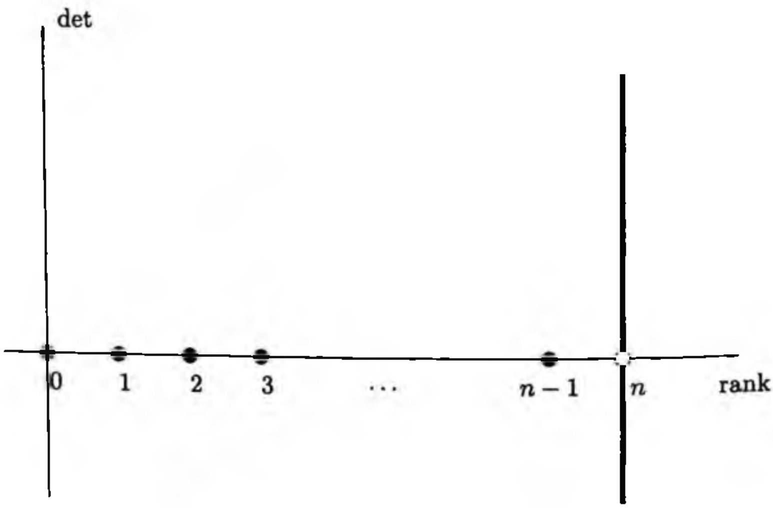


Duality!

Chapter 10

Determinants

Multilinearity appears naturally with volume considerations in vector spaces. A scalar—the determinant—measures the *volume-amplification factor* produced by an operator in a finite-dimensional vector space. This volume-amplification factor is 0 precisely when the operator is singular (not maximal rank), whence its importance for the computation of eigenvalues. Here is a picture illustrating the possible values for the rank and the determinant of an operator.



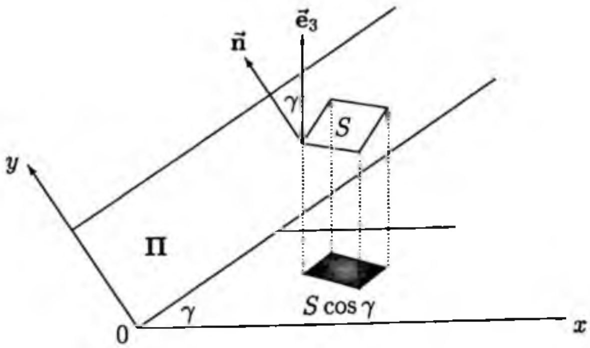
(rank, det) for an $n \times n$ matrix

10.1 From Space Geometry to Determinants

10.1.1 Areas in $\mathbf{R}^3$

The relation between the area of a surface and the area of its orthogonal projection onto a plane is similar to the one given in Sec. 7.1.1 for lengths. A picture shows that the ratio is given by the cosine of the angle of the two planes. Indeed, one linear dimension is preserved by projection, while the orthogonal linear dimension is shrunk by a factor equal to this cosine.

Let us consider any region of a plane $\mathbf{\Pi} \subset \mathbf{R}^3$ having area S, and its vertical projection onto the xy-plane having area $S_{xy} = S \cos\gamma$ (see picture).



Now, the cosine of the angle between $\mathbf{\Pi}$ and the horizontal plane is given by the inner product of the two normals $\vec{\mathbf{n}}$ and $\vec{\mathbf{e}}_3$ to these planes (the unit normal $\vec{\mathbf{n}}$ is defined up to a sign), so that

$$S_{xy} = S\,\vec{\mathbf{n}} \cdot \vec{\mathbf{e}}_3.$$

If we define a vector

$$\vec{\mathbf{S}} = S\,\vec{\mathbf{n}},$$

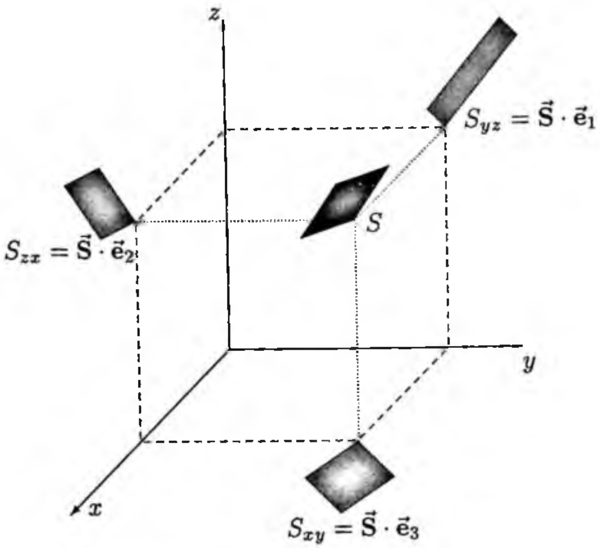
we get

$$S_{xy} = \vec{\mathbf{S}} \cdot \vec{\mathbf{e}}_3,$$

namely:

The area of the projection S_{xy} is the third component of $\vec{\mathbf{S}}$.

Since the definition of the vector $\vec{\mathbf{S}}$ is independent from the direction of projection (again up to a sign), we get similarly the areas of the horizontal projections onto the yz– and zx–planes



The three components S_{yz}, S_{zx}, and S_{xy} of the vector $\vec{S}$ in the orthonormal basis $(\vec{e}_i)$ are *up to signs*, the areas of the projections of the planar region considered. By the Pythagorean theorem

$$S^2 = S_{yz}^2 + S_{zx}^2 + S_{xy}^2.$$

Application. Let us compute the area of a triangle ABC having its three vertices on the axes, say

$$\overrightarrow{OA} = a\,\vec{e}_1, \quad \overrightarrow{OB} = b\,\vec{e}_2, \quad \overrightarrow{OC} = c\,\vec{e}_3.$$

The areas of its orthogonal projections are

$$\text{area } OBC = \tfrac{1}{2}bc, \quad \text{area } OCA = \tfrac{1}{2}ca, \quad \text{area } OAB = \tfrac{1}{2}ab.$$

and thus

$$S^2 = \tfrac{1}{4}(b^2c^2 + c^2a^2 + a^2b^2).$$

Hence the area of the triangle ABC is

$$S = \tfrac{1}{2}\sqrt{b^2c^2 + c^2a^2 + a^2b^2}.$$

Comment. We can write the unit vector $\vec{n}$ orthogonal to Π as

$$\vec{n} = \cos\alpha\,\vec{e}_1 + \cos\beta\,\vec{e}_2 + \cos\gamma\,\vec{e}_3.$$

Hence

$$\cos^2\alpha + \cos^2\beta + \cos^2\gamma = 1.$$

Here α, β, and γ (resp. $\cos\alpha$, $\cos\beta$, and $\cos\gamma$) are the *direction angles* (resp. *direction cosines*) of $\vec{n}$ with the coordinate axes in $\mathbf{R}^3$. The preceding relation is a generalization of

$$\cos^2\alpha + \sin^2\alpha = 1.$$

Indeed, if we consider a unit vector

$$\vec{n} = \cos\alpha\,\vec{e}_1 + \cos\beta\,\vec{e}_2$$

in the plane $\mathbf{R}^2$ (α and β are the direction angles with the coordinate axes in $\mathbf{R}^2$), we also have

$$\cos^2\alpha + \cos^2\beta = 1$$

($\cos\beta = \sin\alpha$, since $\beta = \frac{\pi}{2} - \alpha$).

10.1.2 The Cross Product in $\mathbf{R}^3$

As a preliminary observation, consider two vectors $\begin{pmatrix} a \\ b \end{pmatrix}, \begin{pmatrix} c \\ d \end{pmatrix} \in \mathbf{R}^2$. These vectors are proportional when the slopes b/a and d/c are the same (if $a = 0$, then $c = 0$ too). This happens when $ad = bc$, namely when $ad - bc = 0$. Hence we get a *criterion*

$$\begin{pmatrix} a \\ b \end{pmatrix} \text{ and } \begin{pmatrix} c \\ d \end{pmatrix} \text{ independent} \qquad \Longleftrightarrow \qquad ad - bc \neq 0.$$

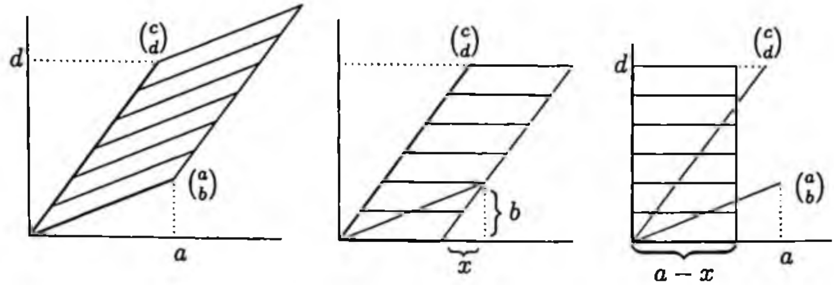
At this point, we may wonder: Has the scalar $ad - bc$ any geometrical meaning? Here is an elementary computation to find it. Introducing polar coordinates, we may write

$$\begin{cases} a = r\cos\varphi \\ b = r\sin\varphi \end{cases} \quad \text{and} \quad \begin{cases} c = s\cos\psi \\ d = s\sin\psi. \end{cases}$$

Hence we find

$$\begin{aligned} ad - bc &= rs(\cos\varphi\sin\psi - \sin\varphi\cos\psi) = rs\sin(\psi - \varphi), \\ &= \pm\,\text{area of the parallelogram generated by } \begin{pmatrix} a \\ b \end{pmatrix} \text{ and } \begin{pmatrix} c \\ d \end{pmatrix}. \end{aligned}$$

Here is a geometrical illustration of this formula.



$b/x = d/c$ $\qquad$ $(a-x)d = ad - bc$

Now let us come to the 3-dimensional case. For two vectors

$$\vec{\mathbf{x}} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad \vec{\mathbf{y}} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \in \mathbf{R}^3,$$

the expressions $x_i y_j - x_j y_i$ $(i < j)$ represent areas (up to sign). As our preliminary observation has shown,

$$x_2y_3 - x_3y_2 = \pm \text{ area of the parallelogram generated by } \begin{pmatrix} x_2 \\ x_3 \end{pmatrix} \text{ and } \begin{pmatrix} y_2 \\ y_3 \end{pmatrix}.$$

This is the area of a projection of the parallelogram generated by $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$: The projection is obtained by forgetting the first component

$$\vec{\mathbf{x}} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \longmapsto \begin{pmatrix} x_2 \\ x_3 \end{pmatrix} \in \mathbf{R}^2.$$

As we have seen in the previous section, the areas of the projections are the components of a vector $\vec{\mathbf{S}}$ which we define as a vector product of $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$.

Definition. *The* cross product *of* $\vec{\mathbf{x}} = (x_i)$ *and* $\vec{\mathbf{y}} = (y_i) \in \mathbf{R}^3$ *is defined in components by*

$$\vec{\mathbf{x}} \wedge \vec{\mathbf{y}} = \begin{pmatrix} x_2y_3 - x_3y_2 \\ x_3y_1 - x_1y_3 \\ x_1y_2 - x_2y_1 \end{pmatrix}.$$

Notice that the index 1 is missing in the first component: $i = 2 < j = 3$ (the index 2 is missing in the second, the index 3 is missing in the third). Instead of the second term $x_1y_3 - x_3y_1$ appearing in $\sum_{i<j}$, we prefer its opposite—having

the same square—for symmetry reasons. In this way, the three components are obtained by circular permutations. Recall now that we proved (Sec. 7.1.2)

$$\|\vec{x}\|^2\|\vec{y}\|^2 - (\vec{x}\cdot\vec{y})^2 = \sum_{i<j}(x_iy_j - x_jy_i)^2.$$

Since

$$\vec{x}\cdot\vec{y} = \|\vec{x}\|\,\|\vec{y}\|\cos\varphi,$$

we recover

$$\begin{aligned}\|\vec{x}\wedge\vec{y}\|^2 &= \|\vec{x}\|^2\|\vec{y}\|^2 - (\vec{x}\cdot\vec{y})^2\\ &= \|\vec{x}\|^2\|\vec{y}\|^2(1-\cos^2\varphi)\\ &= \|\vec{x}\|^2\|\vec{y}\|^2\sin^2\varphi,\\ \|\vec{x}\wedge\vec{y}\| &= \|\vec{x}\|\,\|\vec{y}\|\,|\sin\varphi|.\end{aligned}$$

The norm of $\vec{x}\wedge\vec{y}$ is the area of the parallelogram generated by the vectors $\vec{x}$ and $\vec{y}$, while its components are (up to sign) the areas of its projections.

From the definition, we see immediately that

$$\vec{y}\wedge\vec{x} = -\vec{x}\wedge\vec{y},\quad \vec{x}\wedge\vec{x} = 0.$$

Let us also observe that

$$\begin{aligned}(\vec{x}\wedge\vec{y})\cdot\vec{x} &= (x_2y_3 - x_3y_2)x_1 + (x_3y_1 - x_1y_3)x_2 + (x_1y_2 - x_2y_1)x_3\\ &= y_1(x_3x_2 - x_2x_3) + y_2(-x_3x_1 + x_1x_3) + y_3(x_2x_1 - x_1x_2) = 0,\end{aligned}$$

hence $\vec{x}\wedge\vec{y}\perp\vec{x}$. Since $\vec{x}\wedge\vec{y} = -\vec{y}\wedge\vec{x}$ we also have symmetrically $\vec{x}\wedge\vec{y}\perp\vec{y}$. Hence $\vec{x}\wedge\vec{y}$ is orthogonal to the plane generated by $\vec{x}$ and $\vec{y}$. In particular, if $x_3 = 0$ and $y_3 = 0$ ($\vec{x}$ and $\vec{y}$ horizontal), $\vec{x}\wedge\vec{y} = (x_1y_2 - x_2y_1)\vec{e}_3$ is vertical. Its norm is the absolute value of the third component, equal to the area of the parallelogram in the xy-plane generated by $\vec{x}$ and $\vec{y}$.

Comment. The cross product is only defined in $\mathbf{R}^3$. (We use arrows on vectors as a reminder of this particular situation.) Its components *in the canonical basis* of this space are given by the above formulas, which fix their signs.

10.1.3 The Scalar Triple Product

A combination of the cross product and the inner product leads to an expression for the volumes in $\mathbf{R}^3$. Here appears the *linearity in three variables.*

Theorem. Let $\vec{x}$, $\vec{y}$, and $\vec{z}\in\mathbf{R}^3$. Then

$$(\vec{x}\wedge\vec{y})\cdot\vec{z} = \vec{x}\cdot(\vec{y}\wedge\vec{z})$$

is, up to sign, equal to the volume of the parallelepiped generated by the three vectors $\vec{x}$, $\vec{y}$, and $\vec{z}$. This scalar triple product *vanishes precisely when these vectors are linearly dependent.*

PROOF. We have

$$\begin{aligned}(\vec{\mathbf{x}} \wedge \vec{\mathbf{y}}) \cdot \vec{\mathbf{z}} &= (x_2y_3 - x_3y_2)z_1 + (x_3y_1 - x_1y_3)z_2 + (x_1y_2 - x_2y_1)z_3 \\ &= x_1(y_2z_3 - y_3z_2) + x_2(y_3z_1 - y_1z_3) + x_3(y_1z_2 - y_2z_1) \\ &= \vec{\mathbf{x}} \cdot (\vec{\mathbf{y}} \wedge \vec{\mathbf{z}}).\end{aligned}$$

On the other hand,

$$(\vec{\mathbf{x}} \wedge \vec{\mathbf{y}}) \cdot \vec{\mathbf{z}} = \|\vec{\mathbf{x}} \wedge \vec{\mathbf{y}}\| \, \|\vec{\mathbf{z}}\| \cos\varphi,$$

where φ denotes the angle between $\vec{\mathbf{z}}$ and a normal to the plane generated by $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$. This shows

$$(\vec{\mathbf{x}} \wedge \vec{\mathbf{y}}) \cdot \vec{\mathbf{z}} = \pm \text{ area of parallelogram} \cdot \text{height of parallelepiped},$$

whence the assertion. ∎

Corollary. *The scalar triple product is symmetric with respect to circular permutations of its arguments*

$$\vec{\mathbf{x}} \cdot (\vec{\mathbf{y}} \wedge \vec{\mathbf{z}}) = \vec{\mathbf{y}} \cdot (\vec{\mathbf{z}} \wedge \vec{\mathbf{x}}) = \vec{\mathbf{z}} \cdot (\vec{\mathbf{x}} \wedge \vec{\mathbf{y}}).$$ ∎

Application. Let us consider two distinct straight lines d_1, d_2 in the usual 3-dimensional space. The distance between these lines may be defined as the minimal distance between variable points $P_i \in d_i$. Here is a method to compute it. Take any pair of distinct points A, $B \in d_1$ (resp. C, $D \in d_2$) and consider the vectors

$$\begin{aligned}\vec{\mathbf{d}}_1 &= \overrightarrow{AB} = \overrightarrow{OB} - \overrightarrow{OA}, \\ \vec{\mathbf{d}}_2 &= \overrightarrow{CD} = \overrightarrow{OD} - \overrightarrow{OC},\end{aligned}$$

which furnish the directions of these lines. Now with the vector $\vec{\mathbf{u}} = \overrightarrow{AC}$ linking two points of the lines, the scalar triple product $\vec{\mathbf{u}} \cdot (\vec{\mathbf{d}}_1 \wedge \vec{\mathbf{d}}_2)$ gives the volume of a parallelepiped having basis of area $\|\vec{\mathbf{d}}_1 \wedge \vec{\mathbf{d}}_2\|$ and height equal to the distance of the two lines. Hence we infer

$$\text{distance between } d_1 \text{ and } d_2 = \frac{|\vec{\mathbf{u}} \cdot (\vec{\mathbf{d}}_1 \wedge \vec{\mathbf{d}}_2)|}{\|\vec{\mathbf{d}}_1 \wedge \vec{\mathbf{d}}_2\|}.$$

Summary. (1) Inner products are used for computing angles and lengths, here called norms, and in particular

$$\vec{\mathbf{x}} \perp \vec{\mathbf{y}} \iff \vec{\mathbf{x}} \cdot \vec{\mathbf{y}} = 0.$$

Cross products are used for computing areas, and in particular

$$\vec{\mathbf{x}}, \vec{\mathbf{y}} \text{ proportional} \iff \vec{\mathbf{x}} \wedge \vec{\mathbf{y}} = 0.$$

Scalar triple products are used for computing volumes, and in particular

$$\vec{\mathbf{x}}, \vec{\mathbf{y}}, \text{ and } \vec{\mathbf{z}} \text{ linearly dependent} \iff \vec{\mathbf{x}} \cdot (\vec{\mathbf{y}} \wedge \vec{\mathbf{z}}) = 0.$$

In $\mathbf{R}^3$, we have independence criteria

$$\begin{aligned} \vec{x} \text{ and } \vec{y} \text{ are independent} &\iff \vec{x} \wedge \vec{y} \neq 0, \\ \vec{x},\ \vec{y},\ \text{and } \vec{z} \text{ are independent} &\iff \vec{x} \cdot (\vec{y} \wedge \vec{z}) \neq 0. \end{aligned}$$

Here are more properties of the cross product.

Proposition. *The cross product in* $\mathbf{R}^3$ *satisfies*

(1) LAGRANGE IDENTITY: $(\vec{x} \wedge \vec{y}) \cdot (\vec{u} \wedge \vec{v}) = (\vec{x} \cdot \vec{u})(\vec{y} \cdot \vec{v}) - (\vec{x} \cdot \vec{v})(\vec{y} \cdot \vec{u})$,

(2) GIBBS FORMULA: $\vec{x} \wedge (\vec{y} \wedge \vec{z}) = \vec{y}(\vec{x} \cdot \vec{z}) - \vec{z}(\vec{x} \cdot \vec{y})$,

(3) JACOBI IDENTITY: $\vec{x} \wedge (\vec{y} \wedge \vec{z}) + \vec{y} \wedge (\vec{z} \wedge \vec{x}) + \vec{z} \wedge (\vec{x} \wedge \vec{y}) = 0$.

PROOF. (1) Let us start by the particular case $\vec{v} = \vec{e}_1$ and $\vec{u} = \vec{e}_i$, intending to superpose these particular cases

$$\begin{array}{rrcll} (I) & (\vec{x} \wedge \vec{y}) \cdot (\vec{e}_1 \wedge \vec{e}_1) = & 0 & = x_1 y_1 - x_1 y_1, \\ (II) & (\vec{x} \wedge \vec{y}) \cdot (\vec{e}_2 \wedge \vec{e}_1) = & -3rd \text{ component of } \vec{x} \wedge \vec{y} & = x_2 y_1 - x_1 y_2, \\ (III) & (\vec{x} \wedge \vec{y}) \cdot (\vec{e}_3 \wedge \vec{e}_1) = & 2nd \text{ component of } \vec{x} \wedge \vec{y} & = x_3 y_1 - x_1 y_3, \\ \multicolumn{2}{r}{u_1(I) + u_2(II) + u_3(III) =} & (\vec{x} \cdot \vec{u}) y_1 - x_1 (\vec{y} \cdot \vec{u}). & \end{array}$$

This shows

$$(\vec{x} \wedge \vec{y}) \cdot (\vec{u} \wedge \vec{e}_1) = (\vec{x} \cdot \vec{u})(\vec{y} \cdot \vec{e}_1) - (\vec{x} \cdot \vec{e}_1)(\vec{y} \cdot \vec{u}).$$

Finally, we can still superpose similar results for $\vec{v} = \vec{e}_2$ (resp. $\vec{v} = \vec{e}_3$) and find the announced general formula. Notice that taking in particular $\vec{u} = \vec{x}$ and $\vec{v} = \vec{y}$ we recover a previous formula

$$\|\vec{x} \wedge \vec{y}\|^2 = \|\vec{x}\|^2 \|\vec{y}\|^2 \sin^2 \varphi.$$

(2) Let us determine the components of the double cross product $\vec{x} \wedge (\vec{y} \wedge \vec{z})$ in the canonical orthonormal basis $(\vec{e}_i)$. The first one is

$$\vec{e}_1 \cdot (\vec{x} \wedge (\vec{y} \wedge \vec{z})).$$

This is a scalar triple product, hence invariant under a circular permutation of its arguments, say

$$\vec{e}_1 \cdot (\vec{x} \wedge (\vec{y} \wedge \vec{z})) = (\vec{y} \wedge \vec{z}) \cdot (\vec{e}_1 \wedge \vec{x}).$$

Now, we can use the Lagrange identity (already proved)

$$\begin{aligned} (\vec{y} \wedge \vec{z}) \cdot (\vec{e}_1 \wedge \vec{x}) &= (\vec{y} \cdot \vec{e}_1)(\vec{z} \cdot \vec{x}) - (\vec{y} \cdot \vec{x})(\vec{z} \cdot \vec{e}_1) \\ &= y_1 (\vec{x} \cdot \vec{z}) - z_1 (\vec{x} \cdot \vec{y}). \end{aligned}$$

This is the first component of the announced formula, and the other two are obtained similarly. Observe that the *double cross product* $\vec{x} \wedge (\vec{y} \wedge \vec{z})$ is a vector orthogonal to $\vec{y} \wedge \vec{z}$, hence in the plane generated by $\vec{y}$ and $\vec{z}$:

$$\vec{x} \wedge (\vec{y} \wedge \vec{z}) = a\,\vec{y} + b\,\vec{z}.$$

We have just determined the coefficients a and b of this linear combination.
(3) Let us use the Gibbs formula three times

$$\begin{aligned}
\vec{x} \wedge (\vec{y} \wedge \vec{z}) &= \vec{y}(\vec{x} \cdot \vec{z}) - \vec{z}(\vec{x} \cdot \vec{y}), \\
\vec{y} \wedge (\vec{z} \wedge \vec{x}) &= \vec{z}(\vec{y} \cdot \vec{x}) - \vec{x}(\vec{y} \cdot \vec{z}), \\
\vec{z} \wedge (\vec{x} \wedge \vec{y}) &= \vec{x}(\vec{z} \cdot \vec{y}) - \vec{y}(\vec{z} \cdot \vec{x}).
\end{aligned}$$

Adding these expressions, in the right-hand side terms cancel by pairs, and we get the Jacobi identity. ∎

Remark. The cross product is not associative. In fact, the Jacobi identity

$$\vec{x} \wedge (\vec{y} \wedge \vec{z}) + \vec{y} \wedge (\vec{z} \wedge \vec{x}) + \vec{z} \wedge (\vec{x} \wedge \vec{y}) = 0,$$

can be rewritten

$$\begin{aligned}
\vec{x} \wedge (\vec{y} \wedge \vec{z}) &= -\vec{y} \wedge (\vec{z} \wedge \vec{x}) - \vec{z} \wedge (\vec{x} \wedge \vec{y}) \\
&= \underbrace{(\vec{z} \wedge \vec{x}) \wedge \vec{y}}_{\text{correction!}} + (\vec{x} \wedge \vec{y}) \wedge \vec{z}.
\end{aligned}$$

The deviation from associativity is precisely visible in this form.

The scalar triple product is also denoted by

$$D(\vec{x}, \vec{y}, \vec{z}) = \vec{x} \cdot (\vec{y} \wedge \vec{z}).$$

This function $D : \mathbf{R}^3 \times \mathbf{R}^3 \times \mathbf{R}^3 \to \mathbf{R}$ is linear in each of its variables: We say that it is *trilinear*. Moreover, as we have seen

$$D(\vec{x}, \vec{y}, \vec{z}) = D(\vec{y}, \vec{z}, \vec{x}) = D(\vec{z}, \vec{x}, \vec{y}).$$

Since $\vec{y} \wedge \vec{z} = -\vec{z} \wedge \vec{y}$, we also have

$$D(\vec{x}, \vec{y}, \vec{z}) = -D(\vec{y}, \vec{x}, \vec{z}) = -D(\vec{z}, \vec{y}, \vec{x}) = -D(\vec{x}, \vec{z}, \vec{y}).$$

Our purpose is to generalize these properties to all dimensions $n \geqslant 1$.

10.2 Volume Forms in Vector Spaces

We are going to define a *detector of linear independence* in $\mathbf{R}^n$, similar to the scalar triple product in $\mathbf{R}^3$. This detector will measure the *n-dimensional volume* of a parallelepiped generated by n vectors: It will vanish precisely when the vectors are dependent. When $n = 1$, this volume is (up to sign) a *length*, when $n = 2$ (up to sign) an *area*.

10.2.1 Properties of Volume Forms: Uniqueness

A *volume form* in an n-dimensional space E associates to each family of n elements $\mathbf{v}_i$ in E a scalar $f(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$. This function has to vanish if two vectors of the family are equal, since this situation corresponds to a *flat* or *degenerated parallelepiped*. We also require this function f to be homogeneous of degree 1 in each variable. Referring to our experience of the scalar triple product in the usual space $\mathbf{R}^3$, we postulate homogeneity and additivity, hence linearity in each variable. Thus we shall study *multilinear alternating* functions

$$f : \underbrace{E \times \cdots \times E}_{n \text{ factors}} \longrightarrow \mathbf{R},$$

namely scalar functions in n variables $\mathbf{v}_i \in E$ such that:

- ➤ $f(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ *is linear in each variable*
- ➤ $f(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ *vanishes if two variables have the same value.*

Proposition 1. *A multilinear alternating function f in $m \geqslant 2$ variables $\mathbf{v}_i \in E$, vanishes on all dependent families.*

PROOF. Assume that in the dependent family $(\mathbf{v}_i)_{1 \leqslant i \leqslant n}$, there is a linear dependence relation of the form $\mathbf{v}_1 = \sum_{j>1} a_j \mathbf{v}_j$. Then

$$\begin{aligned} f(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n) &= f\left(\textstyle\sum_{j>1} a_j \mathbf{v}_j, \mathbf{v}_2, \ldots, \mathbf{v}_n\right) \\ &= \sum_{j>1} a_j \underbrace{f(\mathbf{v}_j, \ldots, \mathbf{v}_j, \ldots)}_{=0} = 0. \end{aligned}$$

The situation is similar when any $\mathbf{v}_i$ can be expressed as a function of the other $\mathbf{v}_j$'s. Hence the announced result. ■

This proposition shows that functions of the preceding type with a number of variables m greater than $n = \dim E$ vanish identically.

Proposition 2. *Let $f : E \times E \to \mathbf{R}$ be a bi-additive function. Then the following properties are equivalent:*

$$\begin{array}{ll} (i) & f(\mathbf{v}, \mathbf{v}) = 0 \quad (\mathbf{v} \in E) \\ (ii) & f(\mathbf{v}, \mathbf{w}) = -f(\mathbf{w}, \mathbf{v}) \quad (\mathbf{v}, \mathbf{w} \in E). \end{array}$$

In words: f is *alternating* precisely when it is *skew-symmetric*.

PROOF. $(i) \Rightarrow (ii)$ Let us use the assumption for the sum $\mathbf{v}+\mathbf{w}$ of two arbitrary elements $\mathbf{v}, \mathbf{w} \in E$

$$0 = f(\mathbf{v}+\mathbf{w}, \mathbf{v}+\mathbf{w}) = \underbrace{f(\mathbf{v}, \mathbf{v})}_{=0} + f(\mathbf{v}, \mathbf{w}) + f(\mathbf{w}, \mathbf{v}) + \underbrace{f(\mathbf{w}, \mathbf{w})}_{=0}.$$

We see $f(\mathbf{v}, \mathbf{w}) + f(\mathbf{w}, \mathbf{v}) = 0$.

(*ii*) $\Rightarrow$ (*i*) If f is skew-symmetric, namely

$$f(\mathbf{v},\mathbf{w}) = -f(\mathbf{w},\mathbf{v}) \quad (\mathbf{v},\mathbf{w}\in V),$$

we may simply take $\mathbf{v}=\mathbf{w}$, so that we find

$$f(\mathbf{v},\mathbf{v}) = -f(\mathbf{v},\mathbf{v}), \quad 2f(\mathbf{v},\mathbf{v}) = 0, \quad f(\mathbf{v},\mathbf{v}) = 0. \qquad \blacksquare$$

The following result is obvious.

Proposition 3. If $f: E\times E \to \mathbf{R}$ is linear in the first variable and skew-symmetric, then it is bilinear. $\blacksquare$

The multilinear alternating forms

$$f: \underbrace{E\times\cdots\times E}_{n \text{ factors}} \longrightarrow \mathbf{R} \quad (n=\dim E).$$

are characterized by linearity in the first variable and skew-symmetry in each pair of variables

$$\begin{aligned} f(\mathbf{v}_1+a\mathbf{v}_1',\mathbf{v}_2,\ldots,\mathbf{v}_n) &= f(\mathbf{v}_1,\ldots,\mathbf{v}_n)+af(\mathbf{v}_1',\ldots,\mathbf{v}_n)\\ f(\mathbf{v}_1,\ldots,\mathbf{v}_i,\ldots,\mathbf{v}_j,\ldots,\mathbf{v}_n) &= -f(\mathbf{v}_1,\ldots,\mathbf{v}_j,\ldots,\mathbf{v}_i,\ldots,\mathbf{v}_n). \end{aligned}$$

This set of forms is a vector space $\mathcal{D}=\mathcal{D}(E)$ since sums, and multiples by a scalar, of functions satisfying the preceding conditions, also satisfy them.

Proposition 4. If $f\in\mathcal{D}$ vanishes on a basis, then $f=0$. Two forms $f,g\in\mathcal{D}$ are necessarily proportional.

PROOF. Take a basis $(\mathbf{e}_i)_{1\leqslant i\leqslant n}$ of E. Since any element of the space is a linear combination of the $\mathbf{e}_i$, we may estimate the function f as follows

$$\begin{aligned} f(\Sigma\, a_i\mathbf{e}_i,\Sigma\, b_j\mathbf{e}_j,\Sigma\, c_k\mathbf{e}_k,\ldots) &= \textstyle\sum_i a_i f(\mathbf{e}_i,\Sigma\, b_j\mathbf{e}_j,\Sigma\, c_k\mathbf{e}_k,\ldots)\\ &= \textstyle\sum_{i,j} a_ib_j\, f(\mathbf{e}_i,\mathbf{e}_j,\Sigma\, c_k\mathbf{e}_k,\ldots)\\ &= \textstyle\sum_{i,j,k} a_ib_jc_k\, f(\mathbf{e}_i,\mathbf{e}_j,\mathbf{e}_k,\ldots)\\ &= \textstyle\sum_{i,j,k,\ldots} a_ib_jc_k\cdots f(\mathbf{e}_i,\mathbf{e}_j,\mathbf{e}_k,\ldots). \end{aligned}$$

Now we may forget the monomials corresponding to families of indices $i,j,k,\ldots$ not all distinct since $f(\mathbf{e}_i,\mathbf{e}_j,\mathbf{e}_k,\ldots)=0$ in these cases. Thus we are reduced to summing over families of distinct indices $i,j,k,\ldots$, namely permutations of $1,2,3,\ldots$, say $f(\mathbf{e}_i,\mathbf{e}_j,\mathbf{e}_k,\ldots)=\varepsilon_{ijk\ldots}\, f(\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3,\ldots)$ where all $\varepsilon_{ijk\ldots}$ are ±1. We have found

$$f(\Sigma\, a_i\mathbf{e}_i,\Sigma\, b_j\mathbf{e}_j,\Sigma\, c_k\mathbf{e}_k,\ldots) = \Big(\sum_{i,j,k,\ldots} a_ib_jc_k\cdots\varepsilon_{ijk\ldots}\Big)\, f(\mathbf{e}_1,\mathbf{e}_2,\mathbf{e}_3,\ldots).$$

The knowledge of the single value $f(\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_n)$ determines f completely. If this value is 0, then f vanishes identically. It proves that if f vanishes on a basis (here $(\mathbf{e}_i)$), it vanishes identically. Now we can deduce the second assertion of the proposition. For $f, g \in \mathcal{D}$, consider the linear combination

$$g(\mathbf{e}_1, \ldots, \mathbf{e}_n)f - f(\mathbf{e}_1, \ldots, \mathbf{e}_n)g \in \mathcal{D}.$$

It vanishes on the basis $(\mathbf{e}_i)$ by construction, hence vanishes identically

$$g(\mathbf{e}_1, \ldots, \mathbf{e}_n)f - f(\mathbf{e}_1, \ldots, \mathbf{e}_n)g = 0.$$

This proves the claim $\dim \mathcal{D} \leqslant 1$. ■

Corollary. *The space $\mathcal{D}$ has dimension less than or equal to* 1. ■

Definition. *A* volume form *in E is a nonzero $f \in \mathcal{D}(E)$.*

A volume form f in a finite-dimensional space E of dimension $n \geqslant 1$ is a nonzero multilinear alternating function

$$f : \underbrace{E \times \cdots \times E}_{n \text{ factors}} \longrightarrow \mathbf{R}.$$

It constitutes a *basis* of $\mathcal{D}(E)$: Any $g \in \mathcal{D}(E)$ is a multiple of f. As we have seen, a volume form f cannot vanish on any basis of E. It detects linear independence:

$$\begin{array}{lcl} f(\mathbf{v}_1, \ldots, \mathbf{v}_n) = 0 & \iff & \mathbf{v}_1, \ldots, \mathbf{v}_n \textit{ linearly dependent,} \\ f(\mathbf{v}_1, \ldots, \mathbf{v}_n) \neq 0 & \iff & \mathbf{v}_1, \ldots, \mathbf{v}_n \textit{ linearly independent,} \\ & \iff & \mathbf{v}_1, \ldots, \mathbf{v}_n \textit{ basis of } E. \end{array}$$

In the next section, we shall prove that there is a volume form on any finite-dimensional vector space $E \neq \{0\}$.

If $(\mathbf{e}_i)_{1 \leqslant i \leqslant n}$ is a basis of E, the requirement

$$f(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n) = +1$$

fixes the volume of the unit cube constructed on the chosen basis to be $+1$. It is a *normalization condition.* The formula (found in the proof of Proposition 4)

$$f(\Sigma\, a_i \mathbf{e}_i, \Sigma\, b_j \mathbf{e}_j, \Sigma\, c_k \mathbf{e}_k, \ldots) = \Big(\sum_{i,j,k,\ldots} a_i b_j c_k \cdots \varepsilon_{ijk\ldots} \Big) f(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots)$$

might be used to define a volume form, hence prove its existence. However, if it is easy to see that it defines a multilinear function, it is not so obvious that it is an alternating one. We shall proceed in a more constructive way, furnishing a more effective computing method.

10.2.2 Construction of Volume Forms in $\mathbf{R}^n$

Let us turn to the specific example of the space $E = \mathbf{R}^n$ ($n \geqslant 1$) with the canonical basis $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$.

Definition. The normalized volume form D_n in $\mathbf{R}^n$ is called determinant.

Let us show how to compute this determinant in all positive dimensions. When $n = 1$ we simply take the identity

$$D_1 : \mathbf{R} \longrightarrow \mathbf{R}, \quad a \longmapsto D_1(a) = a.$$

When $n = 2$, a nonzero alternating bilinear function

$$D_2 : \mathbf{R}^2 \times \mathbf{R}^2 \longrightarrow \mathbf{R}$$

is given by

$$D_2(\vec{a}, \vec{b}\,) = D_2\left(\begin{pmatrix} a_1 \\ a_2 \end{pmatrix}, \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}\right) = a_1 b_2 - b_1 a_2.$$

When $n = 3$, the scalar triple product is trilinear alternating, hence a volume form in $\mathbf{R}^3$

$$D_3(\vec{\mathbf{a}}, \vec{\mathbf{b}}, \vec{\mathbf{c}}) = D(\vec{\mathbf{a}}, \vec{\mathbf{b}}, \vec{\mathbf{c}}) = \vec{\mathbf{a}} \cdot (\vec{\mathbf{b}} \wedge \vec{\mathbf{c}}).$$

It is conventional to use the following notation for these determinants (or normalized volume forms)

$$D_2(\vec{a}, \vec{b}\,) = a_1 b_2 - b_1 a_2 = \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix},$$

$$D_3(\vec{\mathbf{a}}, \vec{\mathbf{b}}, \vec{\mathbf{c}}) = \vec{\mathbf{a}} \cdot (\vec{\mathbf{b}} \wedge \vec{\mathbf{c}}) = \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}.$$

(A similar notation for $n = 1$ would be awkward, since it would induce a confusion with the absolute value.)

To give an explicit formula for the normalized volume form $f = D_n$ in $\mathbf{R}^n$, we use induction on the dimension. Here is how to proceed to define $f = D_n$ from $g = D_{n-1}$ ($n \geqslant 2$).

Theorem. Let $\mathbf{v} \mapsto \mathbf{v}' : \mathbf{R}^n \to \mathbf{R}^{n-1}$ denote the linear map that forgets the first component. If g denotes the determinant in $\mathbf{R}^{n-1}$, define $f(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \ldots)$ by

$$a_1\, g(\mathbf{b}', \mathbf{c}', \mathbf{d}', \ldots) - b_1\, g(\mathbf{a}', \mathbf{c}', \mathbf{d}', \ldots) + c_1\, g(\mathbf{a}', \mathbf{b}', \mathbf{d}', \ldots) \mp \cdots$$

Then f is the determinant in $\mathbf{R}^n$.

PROOF. We have to check that the function f defined in the statement is the normalized volume form in $\mathbf{R}^n$. It is obvious that f is linear in each variable and vanishes when two *consecutive* vectors $\mathbf{v}_i = \mathbf{v}_{i+1}$ coincide. Hence this expression changes sign when we permute two consecutive vectors (Proposition

2). In general, in order to permute $\mathbf{v}_i$ and $\mathbf{v}_{i+k}$, we can start by k consecutive permutations of $\mathbf{v}_i$ allowing it to by-pass successively $\mathbf{v}_{i+1}, \ldots, \mathbf{v}_{i+k}$, ending up beyond the last mentioned

$$\mathbf{v}_{i+1}, \ldots, \mathbf{v}_{i+k}, \mathbf{v}_i.$$

Then, $k-1$ permutations of consecutive terms will bring $\mathbf{v}_{i+k}$ in first position. In all, $2k-1$ (odd) changes of sign will occur in the permutation of $\mathbf{v}_i$ and $\mathbf{v}_{i+k}$. This precisely gives a sign change. Finally, for the canonical basis $(\mathbf{e}_i)$ of $\mathbf{R}^n$

$$f(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots) = 1\, g(\mathbf{e}_2', \mathbf{e}_3', \ldots).$$

Note that $\mathbf{e}_2'$, $\mathbf{e}_3', \ldots$ is the canonical basis of $\mathbf{R}^{n-1}$. Since we are assuming that the volume form g is normalized, we have $g(\mathbf{e}_2', \mathbf{e}_3', \ldots) = +1$, and f is also normalized. ■

Example. The inductive definition in dimension 3 (from dimension 2) reads

$$f(\vec{\mathbf{a}}, \vec{\mathbf{b}}, \vec{\mathbf{c}}) = a_1 g\left(\begin{pmatrix} b_2 \\ b_3 \end{pmatrix}, \begin{pmatrix} c_2 \\ c_3 \end{pmatrix}\right) - b_1 g\left(\begin{pmatrix} a_2 \\ a_3 \end{pmatrix}, \begin{pmatrix} c_2 \\ c_3 \end{pmatrix}\right) + c_1 g\left(\begin{pmatrix} a_2 \\ a_3 \end{pmatrix}, \begin{pmatrix} b_2 \\ b_3 \end{pmatrix}\right).$$

With the determinant notation, we obtain the expansion formula

$$\begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} = a_1 \begin{vmatrix} b_2 & c_2 \\ b_3 & c_3 \end{vmatrix} - b_1 \begin{vmatrix} a_2 & c_2 \\ a_3 & c_3 \end{vmatrix} + c_1 \begin{vmatrix} a_2 & b_2 \\ a_3 & b_3 \end{vmatrix},$$

and we find

$$\begin{aligned} D_3(\vec{\mathbf{a}}, \vec{\mathbf{b}}, \vec{\mathbf{c}}) &= a_1(b_2c_3 - b_3c_2) - b_1(a_2c_3 - a_3c_2) + c_1(a_2b_3 - a_3b_2) \\ &= a_1b_2c_3 + a_2b_3c_1 + a_3b_1c_2 - a_1b_3c_2 - a_2b_1c_3 - a_3b_2c_1 \\ &= \vec{\mathbf{a}} \cdot (\vec{\mathbf{b}} \wedge \vec{\mathbf{c}}). \end{aligned}$$

Here is the 4-dimensional case

$$\begin{vmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \\ a_4 & b_4 & c_4 & d_4 \end{vmatrix} = a_1 \begin{vmatrix} b_2 & c_2 & d_2 \\ b_3 & c_3 & d_3 \\ b_4 & c_4 & d_4 \end{vmatrix} - b_1 \begin{vmatrix} a_2 & c_2 & d_2 \\ a_3 & c_3 & d_3 \\ a_4 & c_4 & d_4 \end{vmatrix} + \cdots$$

As each 3×3 determinant is a sum of 6 terms, the expansion of a 4×4 determinant produces $4 \times 6 = 24 = 4!$ monomials.

Observe that we may consider the determinant D_n as a function on $n \times n$ matrices, instead of a function on families of $n$tuples. The expansion of D_n in terms of D_{n-1} is very simple when the first row of the matrix has only one nonzero entry: For example

$$\begin{vmatrix} a_1 & 0 & 0 & \cdots \\ a_2 & b_2 & c_2 & \cdots \\ a_3 & b_3 & c_3 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{vmatrix} = a_1 \begin{vmatrix} b_2 & c_2 & \cdots \\ b_3 & c_3 & \cdots \\ \vdots & \vdots & \ddots \end{vmatrix}.$$

By induction, we immediately obtain the following statement.

Proposition. *The determinant of a lower triangular matrix is the product of the diagonal entries*

$$\begin{vmatrix} a_{11} & 0 & 0 & \dots & 0 \\ a_{21} & a_{22} & 0 & \dots & 0 \\ a_{31} & a_{32} & a_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{vmatrix} = a_{11}a_{22}\cdots a_{nn}. \qquad \blacksquare$$

Moreover, for any volume form $f \in \mathcal{D}(\mathbf{R}^n)$, we have

$$\begin{aligned} f(\dots, \mathbf{e}_i + c\mathbf{e}_j, \dots, \mathbf{e}_j, \dots) &= f(\dots, \mathbf{e}_i, \dots, \mathbf{e}_j, \dots) \\ f(\dots, \mathbf{e}_i, \dots, c\mathbf{e}_i + \mathbf{e}_j, \dots) &= f(\dots, \mathbf{e}_i, \dots, \mathbf{e}_j, \dots) \end{aligned}$$

($1 \leqslant i < j \leqslant n$, c scalar). With $f = D_n$, the first equality is a particular case of the preceding result for lower triangular matrices. The second one concerns its transpose and shows that its determinant is the same. For example if $i = 2 < j = 3$, we have

$$\begin{vmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{vmatrix} = 1 = \begin{vmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 1 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{vmatrix}$$

With the notation of Sec. 3.2.1, this observation concerns the elementary matrices of the form $A = I + cE_{ij}$ $(i \neq j)$, and shows that the determinants of A and its transpose ${}^tA = I + cE_{ji}$ are equal to 1, hence are the same. We shall use this observation in the next section.

10.3 Determinant of an Operator

10.3.1 Volume-Amplification Factor

The scalar triple product in $\mathbf{R}^3$ satisfies

$$D\big(T(\vec{a}), T(\vec{b}), T(\vec{c})\big) = \det(T) D(\vec{a}, \vec{b}, \vec{c}),$$

hence the determinant appears as a volume-amplification factor: The determinant of an operator in any finite-dimensional vector space will similarly be defined as a

volume-amplification factor produced by this operator.

Let us still assume that the finite-dimensional space E is nonzero, so that we can choose a volume form f in E. For any linear operator $T : E \to E$, the composite

$$(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n) \longmapsto (T\mathbf{v}_1, T\mathbf{v}_2, \ldots, T\mathbf{v}_n) \overset{f}{\longmapsto} f(T\mathbf{v}_1, T\mathbf{v}_2, \ldots, T\mathbf{v}_n).$$

is obviously multilinear and skew-symmetric, hence in $\mathcal{D}(E)$, and thus a multiple of the volume form f.

Definition. *The* determinant *of an operator T in E is the proportionality factor in the identity*

$$f(T\mathbf{v}_1, T\mathbf{v}_2, \ldots, T\mathbf{v}_n) = \det(T)\, f(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$$

valid for all families of elements $\mathbf{v}_i$ in E.

Observe that if the family $(\mathbf{v}_i)$ is linearly dependent, so is its image $(T(\mathbf{v}_i))$, and both sides vanish. Since any two volume forms are proportional, this definition of $\det(T)$ is independent from the choice of volume form f. By definition, the determinant of the identity operator is 1.

In $\mathbf{R}^n$, we can choose the normalized volume form D_n, and we find for $\mathbf{v}_j = \mathbf{e}_j$ (canonical basis)

$$D_n(T\mathbf{e}_1, T\mathbf{e}_2, \ldots, T\mathbf{e}_n) = \det(T) D_n(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n) = \det(T).$$

Since $T\mathbf{e}_j = \mathbf{a}_j$ is the jth column of the matrix A of T in the canonical basis

$$\det(T) = D_n(T\mathbf{e}_1, T\mathbf{e}_2, \ldots, T\mathbf{e}_n) = D_n(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n),$$

and

$$\det(T) = \begin{vmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{vmatrix}.$$

This scalar is also written $\det T = \det A$. In particular for the identity operator

$$\det id_E = D_n(\mathbf{e}_1, \ldots, \mathbf{e}_n) = \begin{vmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{vmatrix} = 1$$

Theorem. *The determinant of the composition of two operators in a finite-dimensional space E is the product of their determinants. If A and B are two $n \times n$ matrices,*

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

PROOF. This result is obvious if we refer to the geometrical interpretation of the determinant given above: The volume-amplification factor produced by a composition is the product of the successive volume-amplification factors. It is easy to formalize this proof. Take a basis $(\mathbf{e}_i)$ of E. By definition

$$f((AB)\mathbf{e}_1, (AB)\mathbf{e}_2, \dots, (AB)\mathbf{e}_n) = \det(AB) f(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n).$$

On the other hand, we may take $\mathbf{v}_j = B\mathbf{e}_j$, so that $(AB)\mathbf{e}_j = A(B\mathbf{e}_j) = A\mathbf{v}_j$, and

$$\begin{aligned} f((AB)\mathbf{e}_1, (AB)\mathbf{e}_2, \dots, (AB)\mathbf{e}_n) &= f(A\mathbf{v}_1, A\mathbf{v}_2, \dots, A\mathbf{v}_n) \\ &= \det(A)\, f(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n) \\ &= \det(A)\, f(B\mathbf{e}_1, B\mathbf{e}_2, \dots, B\mathbf{e}_n) \\ &= \det(A)\det(B)\, f(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n). \end{aligned}$$

Since $f \neq 0$, $f(\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n) \neq 0$, and by comparison we obtain

$$\det(AB) = \det(A)\det(B).$$ ∎

Corollary 1. *Let A and B be square matrices of the same size $n \times n$. Then $\det AB = \det BA$. If S is an invertible matrix of the same size, $\det S \neq 0$ and*

$$\det(S^{-1}) = 1/\det S, \quad \det(S^{-1}AS) = \det A.$$

PROOF. The first assertion comes from the multiplication property of determinants

$$\det(AB) = \det A \det B = \det B \det A = \det(BA).$$

If S is invertible, with inverse S^{-1}, we have

$$\det(S^{-1})\det(S) = \det(S^{-1}S) = \det(I) = 1,$$

hence $\det S \neq 0$ and $\det S^{-1} = 1/\det S$. Finally,

$$\det(S^{-1}AS) = \det(AS\,S^{-1}) = \det(A\,I) = \det A.$$

All statements are proved. ∎

Corollary 2. *Let $P \geqslant 0$ be a positive semi-definite matrix. Then $\det P \geqslant 0$.*

PROOF. Since P is symmetric, we may write $P = SDS^{-1}$ where D is a diagonal matrix, having the eigenvalues of P as diagonal entries. By assumption, these eigenvalues are nonnegative, and the determinant of P, equal to the determinant of D, is the product of these nonnegative eigenvalues. ∎

Due to its importance, we formulate and prove the following corollary of the theorem as a separate statement.

Proposition. *A square matrix A of size $n \times n$ is invertible precisely when its determinant is nonzero.*

PROOF. Recall that A is invertible when its rank is maximal. This happens precisely when the columns of A are linearly independent, hence when the determinant of these columns is nonzero. ∎

Applications. (1) We have seen (Sec. 6.4.1) that the Fibonacci sequence may be computed by means of the symmetric matrix $A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$, having determinant -1. We have also proved

$$A^n = \begin{pmatrix} f_{n+1} & f_n \\ f_n & f_{n-1} \end{pmatrix},$$

hence

$$\det A^n = f_{n+1}f_{n-1} - f_n^2 = (-1)^n.$$

In other words

$$f_{n+1}f_{n-1} = f_n^2 + (-1)^n = f_n^2 \pm 1.$$

For example

$$f_7 f_5 = f_6^2 + 1 \ : \ 13 \cdot 5 = 65 = 8^2 + 1$$

$$f_{n+1}f_{n-1} = f_n^2 - 1 = (f_n + 1)(f_n - 1) \quad \text{if } n \text{ is odd.}$$

(2) Consider the plane $\mathbf{R}^2$ as the horizontal subspace of $\mathbf{R}^3$. More precisely, introduce the linear embedding

$$\mathbf{R}^2 \longrightarrow \mathbf{R}^3 : \quad \vec{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \longmapsto \vec{\mathbf{x}} = \begin{pmatrix} x_1 \\ x_2 \\ 0 \end{pmatrix},$$

obtained by adding a third vanishing component. For $\vec{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $\vec{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ in $\mathbf{R}^2$, the definition of the cross product shows that

$$\vec{\mathbf{x}} \wedge \vec{\mathbf{y}} = \begin{pmatrix} 0 \\ 0 \\ x_1 y_2 - x_2 y_1 \end{pmatrix} = \det(\vec{x}, \vec{y})\, \vec{\mathbf{e}}_3.$$

Hence for any 2×2 matrix A, we have

$$\begin{aligned} \overrightarrow{A\vec{x}} \wedge \overrightarrow{A\vec{y}} &= \det(A\vec{x}, A\vec{y})\, \vec{\mathbf{e}}_3 \\ &= \det(A\,(\vec{x}, \vec{y}))\, \vec{\mathbf{e}}_3 \\ &= \det(A) \det(\vec{x}, \vec{y})\, \vec{\mathbf{e}}_3. \end{aligned}$$

Hence the determinant of the 2×2 matrix A is characterized by

$$\overrightarrow{A\vec{x}} \wedge \overrightarrow{A\vec{y}} = \det(A)\, \vec{\mathbf{x}} \wedge \vec{\mathbf{y}}.$$

10.3.2 Determinants and Row Operations

Here is a basic symmetry property of determinants.

Theorem. *For any square matrix A, we have* $\det {}^t\!A = \det A$.

PROOF. If the columns of A are dependent, so are the rows, and both sides vanish. Otherwise $A = E_1E_2 \cdots E_s$ is a product of elementary matrices, and ${}^tA = {}^tE_s \cdots {}^tE_2{}^tE_1$. Since the determinants multiply, it is enough to prove the theorem for each E_i. But two classes of elementary matrices are symmetric, and the third one consists of triangular matrices for which the statement has already been proved (see end of Sec. 10.2.2). ∎

Corollary 1. The determinant of a triangular matrix is the product of its diagonal entries.

PROOF. For lower triangular matrices, this result has already been proved: It is a consequence of the inductive method of computing determinants. For upper-triangular matrices, it now follows from the theorem. ∎

Corollary 2. For any square matrix A, we have

$$\det |A| = |\det A|,$$

where $|A| = ({}^tAA)^{1/2}$ denotes the absolute value of A.

PROOF. By definition $|A|^2 = {}^tAA$ (see Sec. 8.3.3), so that

$$(\det |A|)^2 = \det |A|^2 = \det {}^tAA = \det {}^tA \det A = (\det A)^2,$$

whence $\det |A| = \pm \det A$. Since $|A| \geqslant 0$, we have $\det |A| \geqslant 0$ (10.3.1), and the statement follows. ∎

By definition, the determinant in $\mathbf{R}^n$ is a multilinear skew-symmetric function of n vectors $\mathbf{v}_i \in \mathbf{R}^n$ This implies the following property

$$\begin{aligned} f(\ldots \mathbf{v}_i, \ldots \mathbf{v}_j + c\mathbf{v}_i, \ldots) &= f(\ldots \mathbf{v}_i, \ldots \mathbf{v}_j, \ldots) + c\underbrace{f(\ldots \mathbf{v}_i, \ldots \mathbf{v}_i, \ldots)}_{=0} \\ &= f(\ldots \mathbf{v}_i, \ldots \mathbf{v}_j, \ldots). \end{aligned}$$

Here are the basic rules:

(1) *a determinant changes sign if we exchange two columns*
(2) *if we multiply one column by a scalar a, it is multiplied by a*
(3) *it is unchanged if we add a multiple of one column to another.*

These three properties characterize the behavior of determinants with respect to column operations. Since $\det {}^tA = \det A$, the preceding rules may be reformulated for row operations:

(1)′ *a determinant changes sign if we exchange two rows*
(2)′ *if we multiply one row by a scalar a, it is multiplied by a*
(3)′ *it is unchanged if we add a multiple of one row to another.*

As we know, row operations may always be performed in such a way as to bring the (square) matrix into a triangular shape for which the determinant is the product of the diagonal entries. Hence we may use this method for computing determinants, instead of the inductive method based on the expansion according to the first row.

Finally, note that a combination of the two methods is often used. For example, one may permute the first row with the ith row (hence a change of sign) and then expand the new determinant according to its first line. This corresponds to an expansion of the determinant according to its ith row. In doing so, signs have to be taken care of, and as one may check

$$\begin{vmatrix} a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \\ a_4 & b_4 & c_4 & d_4 \end{vmatrix} = -a_2 \begin{vmatrix} b_1 & c_1 & d_1 \\ b_3 & c_3 & d_3 \\ b_4 & c_4 & d_4 \end{vmatrix} + b_2 \begin{vmatrix} a_1 & c_1 & d_1 \\ a_3 & c_3 & d_3 \\ a_4 & c_4 & d_4 \end{vmatrix} \mp \cdots$$

Rule. To expand a determinant according to an arbitrary row or column, one has to remember that the signs are given by the following *chessboard* of signs

$$\begin{pmatrix} + & - & + & - & \dots \\ - & + & - & + & \dots \\ + & - & + & - & \dots \\ \vdots & & \ddots & \ddots & - \\ & & \dots & - & + \end{pmatrix}.$$

For any fixed row index $1 \leqslant i \leqslant n$, we have

$$\det A = \sum_{1 \leqslant j \leqslant n} (-1)^{i+j} a_{ij} \det A_{ij}$$

where A_{ij} denotes the $(n-1) \times (n-1)$ matrix obtained by erasing the ith row and jth column from A. Since the chessboard of signs has $+$ along its main diagonal, the expansion according to the last row of a determinant is

$$a_{nn} \det A_{nn} - a_{nn-1} \det A_{nn-1} \pm \cdots = \sum_{0 \leqslant j < n} (-1)^j a_{nn-j} \det A_{nn-j}.$$

Corollary. *Let A be a skew-symmetric matrix of size $n \times n$, where n is odd. Then* $\det A = 0$.

PROOF. We have

$$\det A = \det {}^t\!A = \det(-A).$$

To obtain $-A$, each column of A has to be multiplied by -1. Since the determinant is multilinear, $\det(-A) = (-1)^n \det A$. When n is odd, $\det(-A) = -\det A$, and the preceding chain of equalities leads to

$$\det A = \det(-A) = -\det A, \quad 2 \det A = 0, \quad \det A = 0. \qquad \blacksquare$$

Comment. It can be shown that the determinant of a skew-symmetric matrix of *even* size is *nonnegative.* If M is skew-symmetric of size $2n \times 2n$, there is an invertible matrix S (corresponding to a change of basis) such that

$$S^{-1}MS = \begin{pmatrix} 0 & A \\ -A & 0 \end{pmatrix} \quad \text{(block decomposition)},$$

and $\det M = \det S^{-1}MS = (\det A)^2 \geqslant 0$. More precisely, there is a polynomial function of the coefficients of M, called *Pfaffian of* M, denoted $\operatorname{Pf} M$, such that $\operatorname{Pf} M = \det A$. Hence $\det M = (\operatorname{Pf} M)^2 \geqslant 0$ is nonnegative. For example

$$\begin{vmatrix} 0 & a \\ -a & 0 \end{vmatrix} = a^2, \qquad \begin{vmatrix} 0 & a & b & c \\ -a & 0 & d & e \\ -b & -d & 0 & f \\ -c & -e & -f & 0 \end{vmatrix} = (af - be + cd)^2.$$

In the Maxwell electromagnetic theory, the pair consisting of an electric field $\vec{\mathbf{E}}$ and a magnetic field $\vec{\mathbf{B}}$—six components—is replaced by the skew-symmetric matrix

$$\begin{pmatrix} 0 & E_1/c & E_2/c & E_3/c \\ -E_1/c & 0 & -B_3 & B_2 \\ -E_2/c & B_3 & 0 & -B_1 \\ -E_3/c & -B_2 & B_1 & 0 \end{pmatrix}.$$

From the last formula we infer

$$\begin{vmatrix} 0 & E_1/c & E_2/c & E_3/c \\ -E_1/c & 0 & -B_3 & B_2 \\ -E_2/c & B_3 & 0 & -B_1 \\ -E_3/c & -B_2 & B_1 & 0 \end{vmatrix} = (\vec{\mathbf{E}} \cdot \vec{\mathbf{B}})^2/c^2.$$

10.4 Examples of Determinants

There are two methods of computation of determinants. The first one consists in using elementary row (or column) operations—under which the determinant has a known behavior—to bring the matrix into a triangular form. Since the determinant of a triangular matrix is the product of its diagonal entries, we may conclude. For example

$$\begin{vmatrix} 100 & 101 & 102 \\ 101 & 102 & 103 \\ 102 & 103 & 105 \end{vmatrix} = \begin{vmatrix} -1 & -1 & -1 \\ 101 & 102 & 103 \\ 102 & 103 & 105 \end{vmatrix} = \begin{vmatrix} -1 & -1 & -1 \\ 0 & 1 & 2 \\ 0 & 1 & 3 \end{vmatrix} = \begin{vmatrix} -1 & -1 & -1 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{vmatrix} = -1.$$

The second method consists in expanding the determinant according to a row (or a column), to bring it back to a smaller size. This second method is especially successful when one row (or column) has many 0's. In general, a combination of the two methods is used. In some cases, induction or adapted tricks will lead to the final result. Let us give several examples.

10.4.1 Geometric Examples

Example 1. **Area of a Planar Triangle.**

Let us consider the triangle having vertices

$$A = (a_1, a_2), \quad B = (b_1, b_2), \quad \text{and } C = (c_1, c_2).$$

The area of the parallelogram generated by the vectors $\overrightarrow{AB}$ and $\overrightarrow{AC}$ is given by the cross product. Hence

$$2S = \begin{vmatrix} b_1 - a_1 & c_1 - a_2 \\ b_2 - a_2 & c_2 - a_2 \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ a_1 & b_1 - a_1 & c_1 - a_2 \\ a_2 & b_2 - a_2 & c_2 - a_2 \end{vmatrix}.$$

Without changing the determinant, we may add the first column to the second and third, whence

$$S = \frac{1}{2} \begin{vmatrix} 1 & 1 & 1 \\ a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \end{vmatrix}.$$

Example 2. **Volume of a Pyramid.**

Take the pyramid having vertices A, B, C, and D in $\mathbf{R}^3$. The scalar triple product gives the volume of the parallelepiped

$$6V = \overrightarrow{AB} \cdot (\overrightarrow{AC} \wedge \overrightarrow{AD}) = D(\overrightarrow{AB}, \overrightarrow{AC}, \overrightarrow{AD}).$$

Hence

$$V = \frac{1}{6} \begin{vmatrix} b_1 - a_1 & c_1 - a_1 & d_1 - a_1 \\ b_2 - a_2 & c_2 - a_2 & d_2 - a_2 \\ b_3 - a_3 & c_3 - a_3 & d_3 - a_3 \end{vmatrix}.$$

As in the preceding example, we may add a first row $(1\ 0\ 0\ 0)$ and a first column ${}^t(1\ a_1\ a_2\ a_3)$, and eventually, we find

$$V = \frac{1}{6} \begin{vmatrix} 1 & 1 & 1 & 1 \\ a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \end{vmatrix}.$$

Example 3. **Equation of a Circle.**

We are looking for the equation of the circle linking three points P_1, P_2, and P_3 in the plane $\mathbf{R}^2$. Let us recall that the equation of the circle of center (a, b) and radius $r > 0$ is

$$(x - a)^2 + (y - b)^2 = r^2.$$

If we expand this equation, we obtain

$$x^2 + y^2 - 2ax - 2by + (a^2 + b^2 - r^2) = 0.$$

Multiplying by a nonvanishing factor A, we get the equation of the same circle. This shows that the general equation of a circle in the plane has the form

$$A(x^2+y^2)+Bx+Cy+D=0$$

with same coefficient for x^2 and y^2, and no term xy. Now, let us consider the determinant

$$\begin{vmatrix} x^2+y^2 & x & y & 1 \\ x_1^2+y_1^2 & x_1 & y_1 & 1 \\ x_2^2+y_2^2 & x_2 & y_2 & 1 \\ x_3^2+y_3^2 & x_3 & y_3 & 1 \end{vmatrix} = 0.$$

The expansion of this determinant according to its first row leads to the equation

$$(x^2+y^2)\begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix} - xA + yB - C = 0.$$

This is the equation of a circle. We recognize the coefficient of x^2+y^2: It is the double of the area of the triangle $P_1P_2P_3$, hence nonzero precisely when the three given points P_i are not on a line. If we replace (x, y) by the coordinates of a point P_i, we get a determinant with two equal rows, hence with value 0. This proves that the points P_1, P_2, and P_3 lie on the circle.

Remark. The inequality $a^2+b^2 > a^2+b^2-r^2$ is easily translated into the condition $(*)$: $B^2+C^2 > 4AD$ for the equation of the circle. Conversely, any equation $A(x^2+y^2)+Bx+Cy+D=0$, where the coefficients satisfy $(*)$ is the equation of a *real* circle (having a positive radius).

10.4.2 Arithmetic and Algebraic Examples

Example 1. **Fibonacci Numbers.**

Let us consider the following determinants

$$D_n = \begin{vmatrix} 1 & -1 & 0 & 0 & \dots \\ 1 & 1 & -1 & 0 & \dots \\ 0 & 1 & 1 & -1 & \dots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & 1 \end{vmatrix}.$$

For example

$$D_1 = 1, \quad D_2 = \begin{vmatrix} 1 & -1 \\ 1 & 1 \end{vmatrix} = 2, \quad D_3 = \begin{vmatrix} 1 & -1 & 0 \\ 1 & 1 & -1 \\ 0 & 1 & 1 \end{vmatrix} = 3.$$

It is tempting to believe that $D_n = n$ $(n \geqslant 1)$. This is not the case (here is a good place to remember that many sequences have the same first three terms, and indeed, many very natural sequences may have a same beginning: Guessing is no proof!). An expansion according to the first column of D_{n+1} (exercise) eventually shows that $D_{n+1} = D_n + D_{n-1}$. Hence $D_4 = D_3 + D_2 = 3 + 2 = 5$, $D_5 = D_4 + D_3 = 5 + 3 = 8$. We recognize a shifted Fibonacci sequence

$$D_1 = f_2,\ D_2 = f_3,\ D_3 = f_4, \ldots \quad \text{(see Sec. 6.4.1).}$$

This proves

$$D_n = f_{n+1} \quad (n \geqslant 1).$$

Example 2. **Vandermonde Determinants.**

These are determinants having the general form

$$\begin{vmatrix} 1 & a & a^2 & a^3 & \cdots \\ 1 & b & b^2 & b^3 & \cdots \\ 1 & c & c^2 & c^3 & \cdots \\ 1 & d & d^2 & d^3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 & 1 & \cdots \\ a & b & c & d & \cdots \\ a^2 & b^2 & c^2 & d^2 & \cdots \\ a^3 & b^3 & c^3 & d^3 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{vmatrix}.$$

Here are the simplest cases

$$V_1 = 1, \quad V_2 = \begin{vmatrix} 1 & a \\ 1 & b \end{vmatrix} = b - a,$$

$$V_3 = \begin{vmatrix} 1 & a & a^2 \\ 1 & b & b^2 \\ 1 & c & c^2 \end{vmatrix} = \begin{vmatrix} 1 & a & a^2 \\ 0 & b-a & b^2-a^2 \\ 0 & c-a & c^2-a^2 \end{vmatrix} = \begin{vmatrix} b-a & b^2-a^2 \\ c-a & c^2-a^2 \end{vmatrix}$$

$$= (b-a)(c-a)\begin{vmatrix} 1 & b+a \\ 1 & c+a \end{vmatrix} = (b-a)(c-a)(c-b).$$

In general, it is better to write $a = a_1$, $b = a_2$, etc. We intend to prove

$$V_n = \begin{vmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^{n-1} \\ 1 & a_2 & a_2^2 & \cdots & a_2^{n-1} \\ \vdots & \vdots & & & \vdots \\ 1 & a_n & a_n^2 & \cdots & a_n^{n-1} \end{vmatrix} = \prod_{i>j} (a_i - a_j).$$

An astute way of proving this consists in replacing the last row of V_{n+1} by a variable one (replace a_{n+1} by x). The polynomial

$$P(x) = \begin{vmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^n \\ 1 & a_2 & a_2^2 & \cdots & a_2^n \\ \vdots & \vdots & & & \vdots \\ 1 & a_n & a_n^2 & \cdots & a_n^n \\ 1 & x & x^2 & \cdots & x^n \end{vmatrix}$$

has degree $\leqslant n$. If we consider the special values $P(a_j)$ when $1 \leqslant j \leqslant n$, we find a determinant having two equal rows. Hence this polynomial P has the roots $a_1, a_2, \ldots, a_n$, and is proportional to the monic polynomial $\prod_{1 \leqslant j \leqslant n}(x - a_j)$. If we expand the above determinant according to its last row, we find that the leading coefficient of $P(x)$ is the $n \times n$ upper-left sub-determinant V_n:

$$P(x) = V_n \prod_{1 \leqslant j \leqslant n} (x - a_j).$$

Substituting $x = a_{n+1}$, we find

$$V_{n+1} = V_n \cdot P(x)|_{x=a_{n+1}} = V_n \cdot \prod_{1 \leqslant j \leqslant n} (a_{n+1} - a_j).$$

The formula

$$V_n = \prod_{n \geq i > j \geq 1} (a_i - a_j)$$

is now easily seen to hold by induction. A good way to remember which product $\prod_{i<j}$ or $\prod_{i>j}$ to take, is to look at the 2×2 case.

10.4.3 Examples in Calculus

Example 1. **The Mean Value Theorem.**

Let f and g be two real-valued differentiable functions on an interval $[a, b]$. Consider the linear combination of f and g given by

$$\varphi(x) = \begin{vmatrix} f(x) - f(a) & f(b) - f(a) \\ g(x) - g(a) & g(b) - g(a) \end{vmatrix} = \begin{vmatrix} f(a) & f(x) & f(b) \\ g(a) & g(x) & g(b) \\ 1 & 1 & 1 \end{vmatrix}.$$

Obviously $\varphi(a) = \varphi(b) = 0$ (the last determinant has two identical columns in each case). Since φ is differentiable in $[a, b]$, the derivative of φ vanishes at an intermediate point, say

$$\varphi'(\xi) = 0 \quad \text{for some } a < \xi < b.$$

But it is easy to compute the derivative of a 2×2 determinant having variable entries in one column: Simply take the derivative of this column. (The general case is considered below.) Hence

$$\varphi'(\xi) = \begin{vmatrix} f'(\xi) & f(b) - f(a) \\ g'(\xi) & g(b) - g(a) \end{vmatrix} = 0.$$

Assuming $g(b) - g(a) \neq 0$, we infer

$$\frac{f'(\xi)}{g'(\xi)} = \frac{f(b) - f(a)}{g(b) - g(a)}.$$

In particular, me may take the function $g(x) = x$. In this case $g' = 1$ so that

$$\begin{aligned} f'(\xi) &= \frac{f(b)-f(a)}{b-a}, \\ f(b)-f(a) &= f'(\xi)(b-a), \\ f(b) &= f(a) + f'(\xi)(b-a). \end{aligned}$$

Sometimes, this is written slightly differently with $b = a + h$ (h an increment) and $\xi = a + \theta h$:

$$f(a+h) = f(a) + hf'(a+\theta h) \quad \text{for some } 0 < \theta < 1.$$

Note however, that if we apply this weak form of the mean value theorem independently to f and g, we see that

$$\begin{aligned} f(a+h) &= f(a) + hf'(a+\theta_1 h), \\ g(a+h) &= g(a) + hg'(a+\theta_2 h) \end{aligned}$$

for some $\theta_1, \theta_2 \in (0,1)$ (which might be different). The above form shows more precisely that there is a choice $0 < \theta < 1$ for which

$$\frac{f(a+h)-f(a)}{g(a+h)-g(a)} = \frac{f'(a+\theta h)}{g'(a+\theta h)}.$$

This is the *Cauchy Mean Value Theorem.* It is the main step in the derivation of *L'Hôspital's Rule* for computing limits of quotients of differentiable functions.

***Example 2.* The Mean Value Theorem in $\mathbf{R}^3$.**

It is easy to give a generalization of the preceding example. Consider three differentiable functions f, g, and h and form the determinant

$$\Phi(t) = \begin{vmatrix} f(a) & f(t) & f(b) \\ g(a) & g(t) & g(b) \\ h(a) & h(t) & h(b) \end{vmatrix}.$$

As before Φ vanishes at $t = a$ and $t = b$ (the corresponding determinant has two identical columns) and consequently Φ' vanishes in-between, say for some $t = \tau$. This leads to

$$\Phi'(\tau) = \begin{vmatrix} f(a) & f'(\tau) & f(b) \\ g(a) & g'(\tau) & g(b) \\ h(a) & h'(\tau) & h(b) \end{vmatrix} = 0 \quad \text{for some } a < \tau < b.$$

The triple of functions gives a parameterization of a trajectory in the usual space $\mathbf{R}^3$ linking

$$A = (f(a), g(a), h(a)) \text{ (initial point) and } B = (f(b), g(b), h(b)) \text{ (final point).}$$

The derivative gives the *velocity* of the parameterization. Hence the interpretation of the preceding result: There is always an intermediate time τ for which the velocity is parallel to the plane generated by $\overrightarrow{OA}$ and $\overrightarrow{OB}$ (note that here, everything depends on the choice of the origin O). A naive 3-dimensional generalization of the mean value theorem would lead to believing that there is an intermediate τ for which the tangent vector $\Phi'(\tau)$ is proportional to the increment $\overrightarrow{AB} = \overrightarrow{OB} - \overrightarrow{OA}$. This is not true as the following example shows. If the trajectory is a *helix*, the velocity is never parallel to the increment obtained after one whole revolution.

Complement. If a matrix A has differentiable functions as entries, here is how to compute the derivative of its determinant. From the known expansion

$$\det A = D(\Sigma_i a_{i1}\mathbf{e}_i, \Sigma_j a_{j2}\mathbf{e}_j, \ldots) = \sum_{i,j,\ldots} a_{i1}a_{j2}\ldots\varepsilon_{ij\ldots},$$

we find

$$(\det A)' = \sum_{i,j,\ldots} a'_{i1}a_{j2}\cdots\varepsilon_{ij\ldots} + \sum_{i,j,\ldots} a_{i1}a'_{j2}\cdots\varepsilon_{ij\ldots} + \cdots$$

Hence

$$(\det A)' = \begin{vmatrix} a'_{11} & a_{12} & \cdots & a_{1n} \\ a'_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a'_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} + \begin{vmatrix} a_{11} & a'_{12} & \cdots & a_{1n} \\ a_{21} & a'_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a'_{n2} & \cdots & a_{nn} \end{vmatrix} + \cdots$$

Here appears a sum of determinants, each obtained from the given one by derivation of a single column. (By symmetry, one may also derive rows in succession.)

10.4.4 Symbolic Determinants

One can fill a row (or column) of a determinant with vectors instead of scalars. This means that this determinant has to be expanded according to this row. Hence it stands for a linear combination of the vectors in this row: The result is a *vector*. For example

$$\begin{vmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} = \vec{e}_1 \begin{vmatrix} a_2 & a_3 \\ b_2 & b_3 \end{vmatrix} - \vec{e}_2 \begin{vmatrix} a_1 & a_3 \\ b_1 & b_3 \end{vmatrix} + \vec{e}_3 \begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix}.$$

We recognize the definition of the cross product:

$$\begin{vmatrix} \vec{e}_1 & \vec{e}_2 & \vec{e}_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} = \begin{vmatrix} \vec{e}_1 & a_1 & b_1 \\ \vec{e}_2 & a_2 & b_2 \\ \vec{e}_3 & a_3 & b_3 \end{vmatrix} = \vec{a} \wedge \vec{b}.$$

In this determinant representation of the cross product, the orthogonality of $\vec{a} \wedge \vec{b}$ and $\vec{a}$ is visible. Indeed, their dot product is computed as follows

$$\begin{vmatrix} \vec{e}_1 \cdot \vec{a} & \vec{e}_2 \cdot \vec{a} & \vec{e}_3 \cdot \vec{a} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} = \begin{vmatrix} a_1 & a_2 & a_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} = 0.$$

Here is another application of this formalism. Consider three vectors $\vec{a}$, $\vec{b}$, and $\vec{c}$ in the plane $\mathbf{R}^2$ and the symbolic determinant

$$\begin{vmatrix} \vec{a} & \vec{b} & \vec{c} \\ a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \end{vmatrix} = \vec{a}\,A - \vec{b}\,B + \vec{c}\,C.$$

Its first component is

$$(\vec{e}_1 \cdot \vec{a})\,A - (\vec{e}_1 \cdot \vec{b}\,)\,B + (\vec{e}_1 \cdot \vec{c})\,C = a_1 A - b_1 B + c_1 C = \begin{vmatrix} a_1 & b_1 & c_1 \\ a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \end{vmatrix} = 0.$$

The second component vanishes for a similar reason. It proves that

$$\begin{vmatrix} \vec{a} & \vec{b} & \vec{c} \\ a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \end{vmatrix} = \vec{a}\,(b_1c_2 - b_2c_1) - \vec{b}\,(a_1c_2 - a_2c_1) + \vec{c}\,(a_1b_2 - a_2b_1) = \vec{0}.$$

This is a linear dependence relation linking the three given vectors: It is non-trivial if no pair consists of proportional vectors. Similarly,

$$\begin{vmatrix} \vec{a} & \vec{b} & \vec{c} & \vec{d} \\ a_1 & b_1 & c_1 & d_1 \\ a_2 & b_2 & c_2 & d_2 \\ a_3 & b_3 & c_3 & d_3 \end{vmatrix} = \vec{0}$$

gives a linear dependence relation between four vectors in $\mathbf{R}^3$. This linear relation is the *Lagrange identity*

$$\vec{a}\,D(\vec{b}, \vec{c}, \vec{d}) - \vec{b}\,D(\vec{a}, \vec{c}, \vec{d}) + \vec{c}\,D(\vec{a}, \vec{b}, \vec{d}) - \vec{d}\,D(\vec{a}, \vec{b}, \vec{c}) = \vec{0}.$$

The coefficients are the scalar triple products representing volumes of parallelepiped. The proof of such an identity, based solely on the geometric interpretation (volume), seems to be a real challenge. The particular case $\vec{d} = \vec{e}_3$ is interesting

$$\begin{vmatrix} \vec{a} & \vec{b} & \vec{c} & \vec{e}_3 \\ a_1 & b_1 & c_1 & 0 \\ a_2 & b_2 & c_2 & 0 \\ a_3 & b_3 & c_3 & 1 \end{vmatrix} = \vec{0}.$$

An expansion according to the last column gives

$$\begin{vmatrix} \vec{a} & \vec{b} & \vec{c} \\ a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \end{vmatrix} = D(\vec{a}, \vec{b}, \vec{c})\, \vec{e}_3.$$

This is a generalization of the linear dependence relation for three planar vectors. This relation may also be established directly as follows. The first two components of

$$\begin{vmatrix} \vec{a} & \vec{b} & \vec{c} \\ a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \end{vmatrix}$$

vanish (a determinant having two equal rows), while its third component is

$$\begin{vmatrix} a_3 & b_3 & c_3 \\ a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \end{vmatrix} = \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix}.$$

10.5 Appendix

10.5.1 Permutations and Signs

In the expression

$$\det A = D(\Sigma_i a_{i1}\mathbf{e}_i, \Sigma_j a_{j2}\mathbf{e}_j, \ldots) = \sum_{i,j,\ldots} a_{i1} a_{j2} \cdots \varepsilon_{ij\ldots}$$

the signs are given by

$$\varepsilon_{ijk\cdots} = D(\mathbf{e}_i, \mathbf{e}_j, \mathbf{e}_k, \ldots) = -\varepsilon_{jik\cdots}.$$

They are invariant under a circular permutation concerning three indices

$$\varepsilon_{ijk\ell\cdots} = \varepsilon_{jki\ell\cdots} = \varepsilon_{kij\ell\cdots}.$$

Half of the $n!$ permutations on the n integers $\{1, 2, \ldots, n\}$ correspond to the $+$ sign, while the other half correspond to the $-$ sign. The above formula reads

$$\det A = \sum_\sigma \varepsilon_\sigma a_{\sigma(1)1} a_{\sigma(2)2} \cdots$$

where the sum is extended to all permutations σ of $\{1, 2, \ldots, n\}$. The signs ε_σ can also be defined by the following formula

$$\varepsilon_\sigma = \prod_{i>j} (x_{\sigma(i)} - x_{\sigma(j)}) \Big/ \prod_{i>j} (x_i - x_j).$$

10.5.2 More Examples

Example 1. A 4×4 determinant corresponding to a symmetric matrix. To compute

$$\begin{vmatrix} 0 & a & b & c \\ a & 0 & c & b \\ b & c & 0 & a \\ c & b & a & 0 \end{vmatrix},$$

we add the last three rows to the first one

$$\begin{vmatrix} 0 & a & b & c \\ a & 0 & c & b \\ b & c & 0 & a \\ c & b & a & 0 \end{vmatrix} = \begin{vmatrix} a+b+c & a+b+c & a+b+c & a+b+c \\ a & 0 & c & b \\ b & c & 0 & a \\ c & b & a & 0 \end{vmatrix}.$$

Now we have to compute

$$\begin{vmatrix} 1 & 1 & 1 & 1 \\ a & 0 & c & b \\ b & c & 0 & a \\ c & b & a & 0 \end{vmatrix}.$$

From the last three columns, subtract the first one

$$\begin{vmatrix} 1 & 0 & 0 & 0 \\ a & -a & c-a & b-a \\ b & c-b & -b & a-b \\ c & b-c & a-c & -c \end{vmatrix} = \begin{vmatrix} -a & c-a & b-a \\ c-b & -b & a-b \\ b-c & a-c & -c \end{vmatrix}.$$

Add the third row to the first two

$$\begin{vmatrix} -a & c-a & b-a \\ c-b & -b & a-b \\ b-c & a-c & -c \end{vmatrix} = \begin{vmatrix} b-c-a & 0 & b-c-a \\ 0 & a-b-c & a-b-c \\ b-c & a-c & -c \end{vmatrix}$$

$$= (b-c-a)(a-b-c)\begin{vmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ b-c & a-c & -c \end{vmatrix}.$$

Still subtracting the first column from the last, we find

$$(b-c-a)(a-b-c)\begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ b-c & a-c & -b \end{vmatrix}.$$

Finally, the determinant is

$$(a+b+c)(b-c-a)(a-b-c)[-b-a+c] =$$
$$-(a+b+c)(b+c-a)(c+a-b)(a+b-c).$$

It is symmetric in a, b, and c. This determinant is a particular case ($d = 0$) of

$$\begin{vmatrix} d & a & b & c \\ a & d & c & b \\ b & c & d & a \\ c & b & a & d \end{vmatrix} = (d+a+b+c)(d-a+b-c)(d+a-b-c)(d-a-b+c),$$

which will be proved in the next chapter. Here is another generalization

$$\begin{vmatrix} 0 & a & b & c \\ a & 0 & d & e \\ b & d & 0 & f \\ c & e & f & 0 \end{vmatrix} = (af-be-cd)^2 - 4bcde.$$

Example 2. It is surprising that the determinant

$$\begin{vmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & a^2 & b^2 \\ 1 & a^2 & 0 & c^2 \\ 1 & b^2 & c^2 & 0 \end{vmatrix} = -(a+b+c)(b+c-a)(c+a-b)(a+b-c)$$

has the same value as the above one: Subtracting the last column from the second and third ones, we get

$$\begin{vmatrix} 0 & 0 & 0 & 1 \\ 1 & -b^2 & a^2-b^2 & b^2 \\ 1 & a^2-c^2 & -c^2 & c^2 \\ 1 & b^2 & c^2 & 0 \end{vmatrix} = -\begin{vmatrix} 1 & -b^2 & a^2-b^2 \\ 1 & a^2-c^2 & -c^2 \\ 1 & b^2 & c^2 \end{vmatrix}.$$

Subtract now the second row to the first and third ones

$$-\begin{vmatrix} 0 & -b^2-a^2+c^2 & a^2-b^2+c^2 \\ 1 & a^2-c^2 & -c^2 \\ 0 & b^2-a^2+c^2 & 2c^2 \end{vmatrix} = \begin{vmatrix} -b^2-a^2+c^2 & a^2-b^2+c^2 \\ b^2-a^2+c^2 & 2c^2 \end{vmatrix}.$$

Still subtract the second row from the first one

$$\begin{vmatrix} -b^2-a^2+c^2 & a^2-b^2+c^2 \\ b^2-a^2+c^2 & 2c^2 \end{vmatrix} = \begin{vmatrix} -2b^2 & a^2-b^2-c^2 \\ b^2-a^2+c^2 & 2c^2 \end{vmatrix}.$$

We have found

$$\begin{aligned} -4b^2c^2 + (a^2-b^2-c^2)^2 &= [a^2-b^2-c^2+2bc][a^2-b^2-c^2-2bc] \\ &= \left[a^2-(b-c)^2\right]\left[a^2-(b+c)^2\right] \\ &= (a+b-c)(a-b+c)(a-b-c)(a+b+c). \end{aligned}$$

Example 3. The preceding determinant is a particular case of the following one

$$\begin{vmatrix} 0 & d^2 & e^2 & f^2 \\ d^2 & 0 & a^2 & b^2 \\ e^2 & a^2 & 0 & c^2 \\ f^2 & b^2 & c^2 & 0 \end{vmatrix} =$$

$$-(af + be + cd)(be + cd - af)(cd + af - be)(af + be - cd).$$

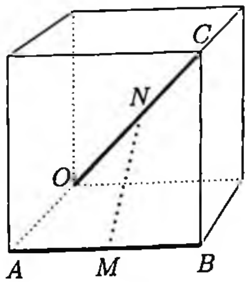
When a, b, c, and d are the lengths of a planar quadrilateral, and e, f its diagonals, we have the *Ptolemy's inequality* (7.2.4)

$$af \leqslant be + cd.$$

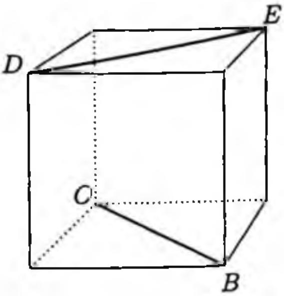
The equality case happens precisely when the quadrilateral can be inscribed in a circle: *Ptolemy's theorem.* In this case, the determinant vanishes. (The triangle inequality only gives $a \leqslant b+c$, a strict inequality as long as the triangle is non-degenerated.)

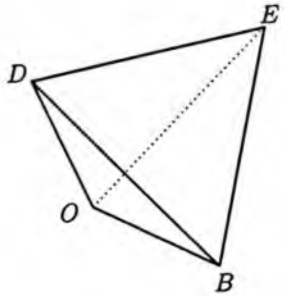
10.6 Exercises

1. Compute the distance between the diagonal and a disjoint edge in the unit cube. Prove that this distance is equal to the length of MN (see picture). Hence OC and MN are orthogonal (check it with the dot product).



2. (*a*) Compute the volume of the regular tetrahedron having edges of unit length. Hint: Consider the regular tetrahedron inscribed in a cube as in the following picture.





(*b*) Compute the volume of the regular octahedron having edges of unit length.

3. What is the determinant of $S = \begin{pmatrix} \lambda & \mu \\ 1 & 1 \end{pmatrix}$. (This matrix was used in Sec. 6.4.1 for the diagonalization of the Fibonacci matrix.)

4. Let $\vec{a}$, $\vec{b}$, and $\vec{c} \in \mathbf{R}^3$ be linearly independent. (*a*) Prove that $\vec{b} \wedge \vec{c}$, $\vec{c} \wedge \vec{a}$, and $\vec{a} \wedge \vec{b}$ are also linearly independent.
(*b*) What is the volume of the parallelepiped generated by $\vec{b} \wedge \vec{c}$, $\vec{c} \wedge \vec{a}$, and $\vec{a} \wedge \vec{b}$? (as a function of the volume of the parallelepiped generated by $\vec{a}$, $\vec{b}$, and $\vec{c}$)?

5. (*a*) Observe that 13 divides 299, 468, and 741. Deduce from this observation that the determinant

$$\begin{vmatrix} 2 & 9 & 9 \\ 4 & 6 & 8 \\ 7 & 4 & 1 \end{vmatrix}$$

is also divisible by 13.
(*b*) From the fact that 17 divides 204, 527, and 255, conclude that 17 also divides the determinant

$$\begin{vmatrix} 2 & 0 & 4 \\ 5 & 2 & 7 \\ 2 & 5 & 5 \end{vmatrix}.$$

6. Compute the determinant

$$A = \begin{vmatrix} 1000 & 1000 & 1000 & 1000 \\ 998 & 999 & 1000 & 1000 \\ 996 & 998 & 1000 & 1000 \\ 1000 & 1000 & 999 & 998 \end{vmatrix}.$$

7. Prove

$$\begin{vmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & a^2 & b^2 \\ 1 & a^2 & 1 & c^2 \\ 1 & b^2 & c^2 & 1 \end{vmatrix} = 2(a^2-1)(b^2-1)(c^2-1).$$

8. (*a*) Establish the Gibbs formula by the following method. First show that the matrix $L_{\vec{x}}$ of the linear map $\vec{u} \mapsto \vec{x} \wedge \vec{u}$ is

$$L_{\vec{x}} = \begin{pmatrix} 0 & -x_3 & \boxed{x_2} \\ \boxed{x_3} & 0 & -x_1 \\ -x_2 & \boxed{x_1} & 0 \end{pmatrix}.$$

Then observe that

$$\vec{x} \wedge (\vec{y} \wedge \vec{z}) = L_{\vec{x}}(L_{\vec{y}}(\vec{z})).$$

(*b*) Show

$$L_{\vec{x}}L_{\vec{y}} - L_{\vec{y}}L_{\vec{x}} = L_{\vec{x}\wedge\vec{y}}.$$

(*c*) Use the Gibbs formula in two different ways to compute the double cross product

$$(\vec{a} \wedge \vec{b}) \wedge (\vec{c} \wedge \vec{d}).$$

Deduce the Lagrange identity by comparison.

9. Define constants ε_{ijk} by $\vec{e}_i \wedge \vec{e}_j = \sum_k \varepsilon_{ijk}\vec{e}_k$. The Lagrange identity leads to

$$\begin{aligned}(\vec{e}_i \wedge \vec{e}_j)(\vec{e}_n \wedge \vec{e}_m) &= \delta_{im}\delta_{jn} - \delta_{in}\delta_{jm},\\ \sum_k \varepsilon_{ijk}\vec{e}_k \cdot \sum_p \varepsilon_{mnp}\vec{e}_p &= \sum_k \varepsilon_{ijk}\varepsilon_{mnk}\\ &= \delta_{im}\delta_{jn} - \delta_{in}\delta_{jm}.\end{aligned}$$

As a consequence, show that the family of ε's satisfies 81 identities!

10. Check the Lagrange identity in $\mathbf{R}^3$

$$(\vec{a} \wedge \vec{b}) \cdot (\vec{c} \wedge \vec{d}) = \begin{vmatrix} \vec{a}\cdot\vec{c} & \vec{a}\cdot\vec{d} \\ \vec{b}\cdot\vec{c} & \vec{b}\cdot\vec{d} \end{vmatrix} = \det \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{pmatrix} \begin{pmatrix} c_1 & d_1 \\ c_2 & d_2 \\ c_3 & d_3 \end{pmatrix}$$

in components (namely without considering first the particular cases $\vec{c} = \vec{e}_i$ and $\vec{d} = \vec{e}_j$).

11. Give a method for finding a linear relation between five vectors in $\mathbf{R}^4$.

12. Let $V = \mathbf{R}^4$ and consider the map

$$f : V \times V \times V \to V,$$

defined by

$$f(\mathbf{a}, \mathbf{b}, \mathbf{c}) = \begin{vmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 & \mathbf{e}_4 \\ a_1 & a_2 & a_3 & a_4 \\ b_1 & b_2 & b_3 & b_4 \\ c_1 & c_2 & c_3 & c_4 \end{vmatrix}.$$

Prove that f is trilinear, alternating (skew-symmetric). Show that the vector $f(\mathbf{a}, \mathbf{b}, \mathbf{c})$ is orthogonal to $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$: It is a kind of ternary cross product. Each component of $f(\mathbf{a}, \mathbf{b}, \mathbf{c})$ is the volume of a parallelepiped (obtained by a projection on a suitable 3-dimensional coordinate subspace of V).

13. Compute the determinants of the following magic squares

4	9	2
3	5	7
8	1	6

,

1	15	14	4
12	6	7	9
8	10	11	5
13	3	2	16

,

10	18	1	14	22
4	12	25	8	16
23	6	19	2	15
17	5	13	21	9
11	24	7	20	3

.

(In a magic square containing the first n^2 integers, the sum of each row and column is $s = \frac{n}{2}(n^2+1)$.)

14. Compute the following determinant

$$\begin{vmatrix} a & 0 & 0 & 0 & 0 & 0 \\ a' & b & b' & b'' & b''' & * \\ a'' & 0 & c & 0 & 0 & 0 \\ a''' & 0 & c' & d & d' & d'' \\ * & 0 & c'' & 0 & e & 0 \\ * & 0 & c''' & 0 & e' & f \end{vmatrix}.$$

15. Compute the following $n \times n$ determinants by induction on n

$$D_n = \begin{vmatrix} 2 & 1 & 0 & \cdots & 0 \\ 1 & 2 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & & \cdots & 1 & 2 \end{vmatrix}, \quad D'_n = \begin{vmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & & \cdots & -1 & 2 \end{vmatrix}.$$

16. Compute the determinants

$$D_n = \begin{vmatrix} 2 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 2 \end{vmatrix}.$$

17. (*a*) Compute

$$\begin{vmatrix} -1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{vmatrix}, \quad \begin{vmatrix} -1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & 1 & -1 \end{vmatrix}, \quad \begin{vmatrix} -1 & -1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{vmatrix}.$$

(*b*) Show that the determinant of an $n \times n$ matrix $A = (a_{ij})$ where all coefficients $a_{ij} = \pm 1$ is divisible by 2^{n-1}.

18. Compute the determinant of the following $2n \times 2n$ matrix

$$\begin{pmatrix} a & 0 & \cdots & \cdots & 0 & b \\ 0 & a & & & b & 0 \\ \vdots & & \ddots & \iddots & & \vdots \\ \vdots & & \iddots & \ddots & & \vdots \\ 0 & b & & & a & 0 \\ b & 0 & \cdots & \cdots & 0 & a \end{pmatrix},$$

using the following methods: (*a*) Use row operations, (*b*) Expand it according to its first row (and use induction), (*c*) Change of basis, using $\mathbf{e}_1, \mathbf{e}_{2n}, \mathbf{e}_2, \mathbf{e}_{2n-1}, \ldots, \mathbf{e}_n, \mathbf{e}_{n+1}$ (as in exercise 2 of Chapter 4).

19. Compute the determinant of the symmetric matrix

$$A_4 = \begin{pmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{pmatrix}.$$

Use the following scheme for the generalization to size $n \times n$: Since A_n is diagonalizable, its determinant is the product of its eigenvalues (accounting for their geometric multiplicities).
Let M_n denote the $n \times n$ matrix having all entries equal to 1. Observe that $\ker M_n$ is an eigenspace of $\alpha I_n + \beta M_n$ with respect to the eigenvalue $\lambda_1 = \alpha$. The other eigenvalue of $\det(\alpha I_n + \beta M_n)$ is easily determined using the trace of this matrix, or observe that the row sums are the same in $\alpha I_n + \beta M_n$, hence we can guess an eigenvector of these matrices.

20. Consider the following $n \times n$ determinant

$$D_n(x) = \begin{vmatrix} 1+x & 1 & \cdots & 1 \\ 1 & 1+x & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1+x \end{vmatrix}.$$

(*a*) Compute $D_1(x)$, $D_2(x)$, and $D_3(x)$.
(*b*) Prove by induction that the derivative of $D_n(x)$ satisfies $D'_n = nD_{n-1}$.
(*c*) Observe that $D_n(0) = 0$. Deduce the value of $D_n(x)$.

21. Give the value of the determinant of

$$\begin{pmatrix} 1 & \cos a & \cos 2a \\ 1 & \cos b & \cos 2b \\ 1 & \cos c & \cos 2c \end{pmatrix}$$

as a product of three terms.

22. Consider square matrices of even size $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$, with blocks of size $n \times n$.

(*a*) Observe that in general

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} \neq \det A \det D - \det B \det C.$$

For this purpose, you may consider the matrices $\begin{pmatrix} I_n & I_n \\ I_n & aI_n \end{pmatrix}$.

(*b*) Observe that in general

$$\det \begin{pmatrix} A & B \\ C & D \end{pmatrix} \neq \det(AD - BC)$$

(nevertheless, it can be shown that the equality holds when $CD = DC$).
(*c*) Establish

$$\det\begin{pmatrix} A & B \\ I_n & I_n \end{pmatrix} = \det(A - B).$$

When D is invertible, it can be shown quite generally that

$$\det\begin{pmatrix} A & B \\ C & D \end{pmatrix} = \det(AD - BD^{-1}CD).$$

23. Let D_n denote the determinant of the following $n \times n$ matrix

$$\begin{pmatrix} a+b & ab & 0 & \dots & \dots & 0 \\ 1 & a+b & ab & 0 & \dots & 0 \\ 0 & 1 & a+b & ab & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & 1 & a+b & ab \\ 0 & 0 & \dots & & 1 & a+b \end{pmatrix}.$$

Show that

$$D_n = (a+b)D_{n-1} - abD_{n-2} \quad (n \geqslant 3),$$

and deduce the value of D_n for all integers n.

24. Compute the determinant

$$\begin{vmatrix} 1+a_1 & 1 & 1 & \dots & 1 \\ 1 & 1+a_2 & 1 & & \\ 1 & 1 & 1+a_3 & & \vdots \\ \vdots & & & \ddots & 1 \\ 1 & \dots & & 1 & 1+a_n \end{vmatrix}.$$

25. (Gram determinants) Let $\mathbf{u}_1, \dots, \mathbf{u}_n, \mathbf{v}_1, \dots, \mathbf{v}_n$ be elements of a Euclidean space V, and let $(\mathbf{e}_k)$ be an orthonormal basis of V, with

$$\mathbf{u}_i = \sum_k a_{ik}\mathbf{e}_k, \quad \mathbf{v}_j = \sum_\ell b_{j\ell}\mathbf{e}_\ell.$$

Show

$$\begin{vmatrix} (\mathbf{u}_1 \mid \mathbf{v}_1) & \cdots & (\mathbf{u}_1 \mid \mathbf{v}_n) \\ \vdots & & \vdots \\ (\mathbf{u}_n \mid \mathbf{v}_1) & \cdots & (\mathbf{u}_n \mid \mathbf{v}_n) \end{vmatrix} = \det A \ \det B,$$

where $A = (a_{ik})$ and $B = (b_{j\ell})$ (see exercise 12 of Chapter 7). Formulate this equality in the usual space $\mathbf{R}^3$, with its canonical basis.

26. Let

$$U_n = U_n(x) = \begin{vmatrix} 2x & 1 & 0 & \dots & & \\ 1 & 2x & 1 & 0 & \dots & \\ 0 & 1 & 2x & 1 & & \dots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \end{vmatrix} \qquad (\text{size } n \times n),$$

so that

$$U_0 = 1, \quad U_1 = 2x, \quad U_2 = 4x^2 - 1.$$

Prove

$$U_{n+1} = 2xU_n - U_{n-1} \quad (n \geqslant 1).$$

Show $U_n(1) = n + 1$ (compare with exercise 4 of Chapter 4). Establish

$$U_n(\cos\varphi) = \frac{\sin(n+1)\varphi}{\sin\varphi} \quad (n \geqslant 0)$$

(Chebyshev polynomials of the second kind).

27. (Smith determinants). Let $a_{ij} = \gcd(i, j)$ denote the greatest common divisor of the two integers i, j. Consider the square matrix $A = A_n = (a_{ij})_{1\leqslant i,j\leqslant n}$ of size $n \times n$. (*a*) Write down explicitly these matrices for $n = 2, 3, 4$, and compute their determinants.
(*b*) Show that $A = TD\,{}^tT$ where T is triangular with entries 0, 1 and $D = \mathrm{diag}(\varphi(1), \ldots, \varphi(n))$ is diagonal. Recall that the Euler φ-function is defined by

$\varphi(k)$ = number of integers in the range $1, \ldots, k$ which are prime to k.

Conclude that

$$\det A_n = \varphi(1)\cdots\varphi(n) = \prod_{1\leqslant k\leqslant n} \varphi(k).$$

Compare with the results obtained under (*a*).

28. Explain the following apparent paradox. The homogeneous system for two complex variables x, y:

$$\begin{cases} (a+ib)x + (c+id)y = 0 \\ (a'+ib')x + (c'+id')y = 0, \end{cases} \tag{HS}$$

has a nontrivial solution exactly when

$$\begin{vmatrix} a+ib & c+id \\ a'+ib' & c'+id' \end{vmatrix} = 0.$$

This complex condition leads to *two real conditions*

$$\begin{cases} ac' - a'c = bd' - b'd \\ ad' + bc' = a'd + b'c. \end{cases}$$

However, writing $x = x_1 + ix_2$, $y = y_1 + iy_2$, and separating real and imaginary parts, the homogeneous system (HS) is equivalent to

$$\begin{cases} ax_1 - bx_2 + cy_1 - dy_2 = 0 \\ bx_1 + ax_2 + dy_1 + cy_2 = 0 \\ a'x_1 - b'x_2 + c'y_1 - d'y_2 = 0 \\ b'x_1 + a'x_2 + d'y_1 + c'y_2 = 0. \end{cases} \tag{$HS)_R$}$$

This system has a nontrivial solution when the *single real condition*

$$\begin{vmatrix} a & -b & c & -d \\ b & a & d & c \\ a' & -b' & c' & -d' \\ b' & a' & d' & c' \end{vmatrix} = 0$$

is satisfied. How is this possible?

Notes

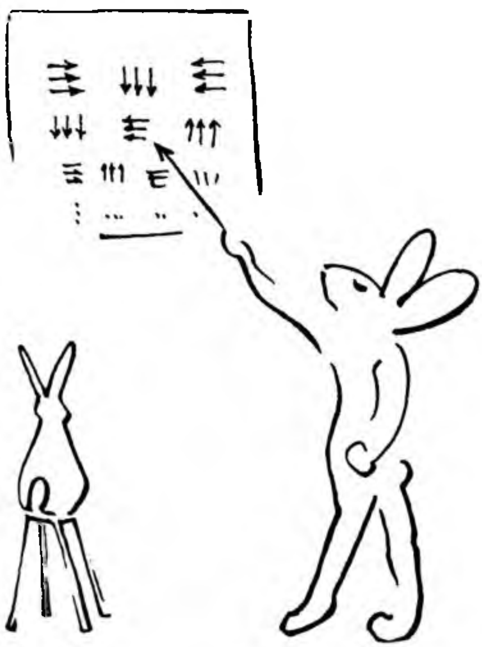
The independence of projections is well illustrated in the frontispiece of the book by Douglas R. Hofstadter:

Gödel-Escher-Bach, (Basic Books 1979).

In Scientific American (August 1991, p.90) the same picture is interpreted as a "Digital Sundial."

Keywords for Web Search

Markus Raetz (artist who plays with the independence of projections).



Checking that he knows how to expand determinants!

Chapter 11

Applications

Let T be an operator in a *finite-dimensional* real vector space $E \neq \{0\}$. Recall that an eigenvector of T is a nonzero $\mathbf{v} \in V$ such that $T\mathbf{v}$ is proportional to $\mathbf{v}$, say $T\mathbf{v} = \lambda\mathbf{v}$. Thus we are interested in finding pairs $(\lambda, \mathbf{v})$ satisfying

$$(T - \lambda I)\mathbf{v} = 0 \text{ and } \mathbf{v} \neq 0.$$

In the *geometric theory* (Chapter 6), we were particularly interested in the eigenvectors $\mathbf{v}$. In the present *algebraic theory*, we are more interested in the eigenvalues: The scalars λ such that $\ker(T - \lambda I) \neq \{0\}$. The corresponding eigenvectors are the nonzero elements in this kernel. But $\ker(T - \lambda I) \neq \{0\}$ means that $T - \lambda I$ is not injective, hence not invertible. The determinant precisely detects this situation:

$$\det(T - \lambda I) = 0.$$

An advantage of this purely algebraic characterization of the eigenvalues of T is that it makes no reference to the eigenvectors (also unknown).

11.1 The Characteristic Polynomial

11.1.1 Definition and Basic Properties

As we have just observed, *the eigenvalues of an operator T are the (real) roots of the polynomial*

$$p_T(x) = \det(T - xI).$$

Definition. *The* characteristic polynomial *of an operator T is*

$$p_T(x) = \det(T - xI).$$

Take a basis $(\mathbf{e}_i)_{1 \leqslant i \leqslant n}$ of E. Replace T by its matrix $A = \mathrm{Mat}_{(e)}(T)$, which operates by left multiplication in $\mathbf{R}^n$. Since the determinant of an operator can be computed in matrix form, we have

$$p_T(x) = \det(T - xI) = \det(A - xI).$$

As we know by Sec. 4.4.4, another choice of basis leads to a similar matrix, say $B = S^{-1}AS$ for which

$$\begin{aligned}
\det(B - xI) &= \det(S^{-1}AS - xS^{-1}S) \\
&= \det\left(S^{-1}(A - xI)S\right) \\
&= \det(S^{-1})\det(A - xI)\det S \\
&= \det(A - xI) = p_A(x).
\end{aligned}$$

This is in agreement with the fact that the characteristic polynomial of T can be computed in any matrix description of this operator. All coefficients of p_A are *invariant* under a change of basis. (Note however that the eigenvectors of A and $S^{-1}AS$ are not the same.)

Proposition. *The characteristic polynomial of an $n \times n$ matrix $A = (a_{ij})$ is*

$$p_A(x) = \prod_{1 \leqslant i \leqslant n} (a_{ii} - x) + q_A(x),$$

where q_A is a polynomial of degree less than or equal to $n - 2$.

PROOF. The statement is correct when $n = 1$ (obvious) and $n = 2$:

$$\begin{vmatrix} a_{11} - x & a_{12} \\ a_{21} & a_{22} - x \end{vmatrix} = (a_{11} - x)(a_{22} - x) - \underbrace{a_{12}a_{21}}_{\text{degree } \leqslant 0}.$$

If it is true in dimension $n - 1$, we may establish its truth in dimension n as follows. We expand the determinant of $A - xI$ according to its first row and, using the induction assumption we obtain

$$\det(A - xI) = (a_{11} - x)\Big(\prod_{2 \leqslant i \leqslant n} (a_{ii} - x) + q(x)\Big) + \sum_{2 \leqslant j \leqslant n} (-1)^{j+1} a_{1j} M_{1j}.$$

Here $q(x)$ is a polynomial of degree less than or equal to $n - 3$, and M_{1j} is the determinant of the sub-matrix A_{1j} of A obtained by erasing the first row and the column containing $a_{jj} - x$. Since A_{1j} contains only $n - 2$ entries depending on x, the factor M_{1j} is a polynomial of degree less than or equal to $n - 2$. We find

$$\begin{aligned}
p_T(x) &= (a_{11} - x)\Big(\prod_{2 \leqslant i \leqslant n} (a_{ii} - x) + q(x)\Big) + (\deg \leqslant n - 2) \\
&= (a_{11} - x) \prod_{2 \leqslant i \leqslant n} (a_{ii} - x) + (\deg \leqslant n - 2).
\end{aligned}$$

The induction step is established and the proposition is proved. ■

Theorem. *The characteristic polynomial of a matrix $A \in M_n(\mathbf{R})$ is an nth degree polynomial having the following form*

$$p_A(x) = (-1)^n x^n + (-1)^{n-1}(\operatorname{tr} A)\, x^{n-1} \pm \cdots + \det A.$$

PROOF. The constant term of the polynomial $p_A(x)$ is $p_A(0) = \det A$. The preceding proposition shows that the coefficients of x^n and of x^{n-1} in $p_A(x)$ and in $\prod_{1\leqslant i\leqslant n}(a_{ii} - x)$ are the same. The highest power of x in this product is obviously $(-x)^n = (-1)^n x^n$. Moreover, all monomials of degree $n-1$ in the product are obtained by taking the constant term a_{ii} in one factor, and the term $-x$ in all other ones. Their contribution is

$$(a_{11} + \cdots + a_{nn})(-x)^{n-1} = (-1)^{n-1}(\operatorname{tr} A)\, x^{n-1}.$$

The theorem is established. ■

Corollary. *Any operator T in an odd-dimensional real vector space E has at least a real eigenvalue.*

PROOF. If $n = \dim E$ is odd, the characteristic polynomial of the operator T is

$$p_T(x) = -x^n + (\text{terms of } \deg \leqslant n-1),$$

hence tends to $-\infty$ when $x \to \infty$ while it tends to $+\infty$ when $x \to -\infty$. By the intermediate value theorem (valid for all continuous functions), it has to vanish at some intermediate point, which is necessarily an eigenvalue. ■

11.1.2 Examples

The characteristic polynomial of a 2×2 matrix A is

$$p_A(x) = x^2 - (\operatorname{tr} A)\, x + \det A.$$

Explicitly, the characteristic polynomial of $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, is

$$x^2 - (a+d)x + (ad - bc).$$

The characteristic polynomial of a diagonal matrix $\operatorname{diag}(d_1, \ldots, d_n)$ is

$$(d_1 - x) \cdots (d_n - x).$$

Since a change of basis does not change the characteristic polynomial, we infer that the characteristic polynomial of a diagonalizable matrix is a similar product of degree one terms, where the d_i's are the eigenvalues of the matrix. For example, we have seen in Sec. 6.4.2 that the $n \times n$ matrix

$$M = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 1 \end{pmatrix}$$

is diagonalizable:

- ➢ $\ker M$ has dimension $n-1$ (eigenspace corresponding to $\lambda = 0$)
- ➢ ${}^t(1,\ldots,1)$ is an eigenvector with the eigenvalue $\lambda = n$.

Hence the characteristic polynomial of this matrix is

$$p_M(x) = (-x)^{n-1}(n-x).$$

In the same way, we see that the characteristic polynomial of

$$\alpha M = \begin{pmatrix} \alpha & \alpha & \ldots & \alpha \\ \alpha & \alpha & \ldots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \ldots & \alpha \end{pmatrix}$$

is $p_{\alpha M}(x) = (-x)^{n-1}(n\alpha - x)$. The characteristic polynomial of $\alpha M + \beta I$ is then

$$\begin{aligned} \det(\alpha M + \beta I - xI) &= p_{\alpha M}(\beta - x) \\ &= (\beta - x)^{n-1}(n\alpha + \beta - x). \end{aligned}$$

The characteristic polynomial of

$$A = \begin{pmatrix} a & b & \ldots & b \\ b & a & \ldots & b \\ \vdots & \vdots & \ddots & \vdots \\ b & b & \ldots & a \end{pmatrix} = bM + (a-b)I$$

is thus

$$p_A(x) = (a-b-x)^{n-1}\big(a + (n-1)b - x\big).$$

In particular, for $x = 0$, we get

$$\det A = (a-b)^{n-1}(a + (n-1)b).$$

Specializing to $a = 0$ and $b = 1$, we obtain

$$\begin{vmatrix} 0 & 1 & \ldots & 1 \\ 1 & 0 & \ldots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \ldots & 0 \end{vmatrix} = (-1)^{n-1}(n-1).$$

11.2 The Spectrum of an Operator

11.2.1 Changing the Field of Scalars

Although we have mainly treated vector spaces over the real field $\mathbf{R}$, it is important to consider other cases, and especially vector spaces over the complex field $\mathbf{C}$ (Sec. 3.3.3). It is obvious that the reduction algorithm for the solution of linear systems works over any field of scalars and furnishes solutions (if they exist), in the smallest field containing the coefficients. It is also obvious that the construction of volume forms in $\mathbf{R}^n$ can be performed in the spaces K^n where K is any field of scalars. Similarly, any K-linear map has a determinant which is an element of K. Any $n \times n$ matrix A with coefficients in K defines a K-linear map $K^n \to K^n$, and $\det A \in K$. For $K = \mathbf{R}$, considered as a subfield of $\mathbf{C}$, A defines linear maps

$$\mathbf{R}^n \longrightarrow \mathbf{R}^n \text{ (R-linear)}, \quad \mathbf{C}^n \longrightarrow \mathbf{C}^n \text{ (C-linear)}.$$

They have the same determinant: $\det A \in \mathbf{R} \subset \mathbf{C}$. In particular, the characteristic polynomials of the preceding linear maps are the same.

11.2.2 Roots of the Characteristic Polynomial

Let $T : E \to E$ be an operator in a finite-dimensional *real* vector space $E \neq \{0\}$.

Definition. *The complex roots of the characteristic polynomial of T are called* spectral values. *The set of spectral values is the* spectrum $\sigma = \sigma_T \subset \mathbf{C}$ *of the operator T.*

Theorem. *Let $\lambda \in \sigma$ be a spectral value of T. Then:*

If λ is real, it is an eigenvalue
If $\lambda \notin \mathbf{R}$ there is a 2-dimensional subspace V such that $T(V) = V$.

PROOF. The first assertion is obvious (Chapter 6). In this case, the dimension of the corresponding eigenspace $E_\lambda = \ker(T - \lambda I)$ is the geometric multiplicity $m_\lambda \geqslant 1$ of λ. To prove the second assertion, choose a basis of E and replace E by $\mathbf{R}^n$, and T by its matrix A in this basis of V. Take a complex eigenvector $\mathbf{v} \in \mathbf{C}^n$, with respect to the spectral value $\lambda \notin \mathbf{R}$ of the operator $A : \mathbf{C}^n \to \mathbf{C}^n$ (given by matrix multiplication). Separate the real and imaginary parts: Say

$$\lambda = \lambda' + i\lambda'', \quad \mathbf{v} = \mathbf{v}' + i\mathbf{v}'' \quad (\lambda', \lambda'' \in \mathbf{R};\ \mathbf{v}', \mathbf{v}'' \in \mathbf{R}^n).$$

The complex identity

$$A(\mathbf{v}' + i\mathbf{v}'') = (\lambda' + i\lambda'')(\mathbf{v}' + i\mathbf{v}''),$$

leads to two real equalities

$$\begin{cases} A\mathbf{v}' = \lambda'\mathbf{v}' - \lambda''\mathbf{v}'' \\ A\mathbf{v}'' = \lambda''\mathbf{v}' + \lambda'\mathbf{v}''. \end{cases}$$

This shows that the subspace $V = \mathcal{L}(\mathbf{v}', \mathbf{v}'') \subset \mathbf{R}^n$ generated by the real vectors $\mathbf{v}'$ and $\mathbf{v}''$ is invariant under A. Since λ is not real, $\lambda'' \neq 0$ and $\mathbf{v}$ is not a multiple of a real vector: $\mathbf{v}'$ and $\mathbf{v}''$ are independent. Hence V has dimension 2. ■

Recall now that a square matrix $A = (a_{ij})$ of size $n \times n$ (and real entries) having large diagonal entries is invertible. More precisely, the Gershgorin theorem (Sec. 4.3.3) asserts that if

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \quad (1 \leqslant i \leqslant n),$$

then A is invertible. The same result holds for the complex matrices $A - \lambda I$ ($\lambda \in \mathbf{C}$), and this proves that a spectral value λ of A can occur only if one inequality

$$|a_{ii} - \lambda| \leqslant \sum_{j \neq i} |a_{ij}|$$

holds for some $1 \leqslant i \leqslant n$ (Sec. 6.3.2). Since spectral values of A are eigenvalues of the linear operator $\mathbf{C}^n \to \mathbf{C}^n$ defined by multiplication by A, hence corresponds to an eigenvector $\mathbf{v}$ in $\mathbf{C}^n$, the same estimates as in the real case prove the following result.

***Theorem** (**Gershgorin**). The spectrum of a square matrix $A = (a_{ij})$ is contained in the union of the complex discs*

$$B_i: \quad |z - a_{ii}| \leqslant r_i = \sum_{j \neq i} |a_{ij}|. \qquad \blacksquare$$

Since the field of complex numbers $\mathbf{C}$ is *algebraically closed*, the characteristic polynomial of A is a product of first degree factors:

$$p_A(x) = \prod_{1 \leqslant i \leqslant n} (\lambda_i - x) \quad (\lambda_i \in \mathbf{C})$$

(we have adapted the signs in order that both sides have the same leading coefficient $(-1)^n$). Here, multiple roots correspond to repeated factors.

***Definition.** The* algebraic multiplicity *of a spectral value λ is the highest power of $x - \lambda$ that divides the characteristic polynomial p_T.*

A root λ of a polynomial p has algebraic multiplicity μ when

$$p(\lambda) = p'(\lambda) = \cdots = p^{(\mu-1)}(\lambda) = 0 \text{ but } p^{(\mu)}(\lambda) \neq 0.$$

Comparing

$$p_A(x) = \prod_{1 \leqslant i \leqslant n} (\lambda_i - x) \quad (\lambda_i \in \mathbf{C}).$$

with the general form of the characteristic polynomial

$$p_A(x) = (-1)^n x^n + (-1)^{n-1} (\operatorname{tr} A)\, x^{n-1} \pm \cdots + \det A,$$

we infer

$$\det A = \prod_{1 \leqslant i \leqslant n} \lambda_i, \quad \operatorname{tr} A = \sum_{1 \leqslant i \leqslant n} \lambda_i.$$

If we call μ_λ the algebraic multiplicity of a spectral value λ, grouping together the spectral values λ_i which coincide with one $\lambda \in \sigma$ we find

$$p_A(x) = \prod_{1 \leqslant i \leqslant n} (\lambda_i - x) = \prod_{\lambda \in \sigma} (\lambda - x)^{\mu_\lambda}.$$

Finally, comparing the degrees in

$$p_A(x) = \prod_{\lambda \in \sigma} (\lambda - x)^{\mu_\lambda},$$

we infer

$$n = \deg p_A = \sum_{\lambda \in \sigma} \mu_\lambda.$$

Application. Consider the matrix

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

Since the four row sums are the same, a first eigenvector is given by ${}^t(1\ 1\ 1\ 1)$. Now, considering that the rows are permuted from the first one, we may try eigenvectors having entries ± 1. Eventually, we find that

$$\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix},$$

are eigenvectors with respective eigenvalues

$$a+b+c+d, \quad a-b+c-d, \quad a-b-c+d, \quad a+b-c-d.$$

The trace of A is $4a$: It is the sum of the eigenvalues. The determinant is the product of the eigenvalues

$$\det A = (a+b+c+d)(a-b+c-d)(a-b-c+d)(a+b-c-d).$$

Finally, the characteristic polynomial of A is the product of the $\lambda_i - x$ (it can also be obtained by replacing a by $a - x$ in the determinant)

$$(a+b+c+d-x)(a-b+c-d-x)(a-b-c+d-x)(a+b-c-d-x).$$

Proposition. *The geometric multiplicity of an eigenvalue is less than or equal to its algebraic multiplicity:*

$$m_\lambda = \dim \ker(T - \lambda I) \leqslant \mu_\lambda.$$

PROOF. Take a basis $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m)$ of the eigenspace $V_\lambda = \ker(T - \lambda I) \subset E$, and complete it into a basis of E. In this basis, the matrix A of the operator T consists of four blocks: The first one is λI_m followed by a zero block just below. The matrix of $T - xI$ in the same basis also has four blocks: Here is how it looks like

$$A - xI = \begin{pmatrix} (\lambda - x)I_m & * \\ 0 & * \end{pmatrix}.$$

Expanding according to the first column repeatedly (m times), we find that $p_A(x) = (\lambda - x)^m \cdot (*)$ is divisible by $(\lambda - x)^m$. This proves $\mu \geqslant m$. ■

Corollary 1. A square matrix A can be diagonalized over $\mathbf{R}$ *precisely when $\sigma \subset \mathbf{R}$ and $m_\lambda = \mu_\lambda$ for all $\lambda \in \sigma$.*

PROOF. Indeed, if a single inequality $m_\lambda < \mu_\lambda$ holds, $\sum m_\lambda < \sum \mu_\lambda = n$, so that there is no basis of $\mathbf{R}^n$ consisting of eigenvectors of A. ■

Recall the simplest case (Sec. 6.4.2), systematically considered in elementary textbooks.

Corollary 2. When the characteristic polynomial p_A has n distinct real roots, then the matrix A can be diagonalized. ■

Corollary 3. A triangular matrix $A = (a_{ij})$ having distinct diagonal entries a_{ii}, can be diagonalized.

PROOF. Indeed, if A is triangular, so is $A - xI$ and

$$p_A(x) = \det(A - xI) = \prod_{1 \leqslant i \leqslant n} (a_{ii} - x).$$

Hence the announced result follows from the previous corollary. ■

Remark. This theory can be made with any field of scalars F, if one counts algebraic multiplicities in an algebraic closure of F. But note that geometric multiplicities depend on the base field. For example

$$m_\lambda(\mathbf{R}) \leqslant m_\lambda(\mathbf{C}) \leqslant \mu_\lambda.$$

When the characteristic polynomial of A has n distinct complex roots, then it is possible to diagonalize A in a suitable basis of $\mathbf{C}^n$.

Examples. (a) Consider a matrix $A = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix}$. Its characteristic polynomial is

$$\begin{aligned} x^2 - (\operatorname{tr} A)x + \det A &= x^2 - (2\cos\varphi)x + 1 \\ &= (x - e^{i\varphi})(x - e^{-i\varphi}). \end{aligned}$$

When the angle φ is not an integral multiple of π, the spectral values $e^{\pm i\varphi}$ are distinct, and A can be diagonalized (when φ is an integral multiple of π, A *is* diagonal).

(b) Let N be a nilpotent operator in $\mathbf{R}^n$. The only complex eigenvalue of N is 0, hence the only spectral value of N is 0: $p_N(x) = (-x)^n$. The algebraic multiplicity of this spectral value is n while the geometric multiplicity is $\dim \ker N < n$ if $N \neq 0$ (this implies $n > 1$). Hence N cannot be diagonalized (even over the algebraically closed field $\mathbf{C}$).

(c) When N is nilpotent, the only complex eigenvalue of $I \pm N$ is 1. Hence the operators $I + N$ and $I - N$ have the same characteristic polynomial $(1-x)^n$.

11.2.3 Existence of a Complex Eigenvalue

Since the characteristic polynomial of an operator always has a root in $\mathbf{C}$ (since this field is algebraically closed), we infer that any operator $T : \mathbf{R}^n \to \mathbf{R}^n$ has a *complex* eigenvalue, hence has an eigenvector in $\mathbf{C}^n$. Consequently, the following version of the Schur lemma (Sec. 6.3.4) holds.

***Lemma** (Schur's lemma). Let $(S_i)_{i \in I}$ be an* irreducible *family of operators in a finite-dimensional space E over the complex field $\mathbf{C}$. Then any operator T in E that commutes with all S_i is a multiple of the identity.* ■

Another application of the existence of complex eigenvalues concerns the possibility of finding a complex triangular form for any square matrix.

***Definition.** A square matrix $A = (a_{ij})$ having real or complex coefficients, is said to be* trigonalizable *over the field $\mathbf{R}$ (resp. $\mathbf{C}$), when there exists a real (resp. complex) invertible matrix S (corresponding to a change of basis) such that $S^{-1}AS = T$ is upper-triangular*

$$A = STS^{-1}, \quad T = (t_{ij}) \text{ where } t_{ij} = 0 \text{ for all pairs } i > j.$$

***Proposition.** Any square matrix having real or complex coefficients, is trigonalizable over the field $\mathbf{C}$.*

PROOF. Let $\mathbf{v}$ be an eigenvector of A: $\mathbf{v} \neq 0$ and $A\mathbf{v} = \lambda\mathbf{v}$. There is a basis of $\mathbf{C}^n$ having first element $\mathbf{v}$. In such a basis, the matrix of A is

$$S^{-1}AS = \begin{pmatrix} \lambda & * \cdots * \\ 0 & \\ \vdots & \boxed{A_0} \\ 0 & \end{pmatrix}$$

This observation is the basis for a proof of the proposition by induction on the size of A. There is nothing to prove if A has size 1×1. Assume that the result is true for matrices of size $(n-1) \times (n-1)$. Then if A has size $n \times n$, in the above formula A_0 has size $(n-1) \times (n-1)$, hence there is a basis of $\mathbf{C}^{n-1}$ in which it is upper-triangular, say $S_0^{-1}A_0S_0 = T_0$. Define

$$S_1 = \begin{pmatrix} 1 & 0 \cdots 0 \\ 0 & \\ \vdots & \boxed{S_0} \\ 0 & \end{pmatrix}, \quad S_1^{-1} = \begin{pmatrix} 1 & 0 \cdots 0 \\ 0 & \\ \vdots & \boxed{S_0^{-1}} \\ 0 & \end{pmatrix}.$$

Then, block multiplication shows that

$$S_1^{-1}\begin{pmatrix} \lambda & * \cdots * \\ 0 & \\ \vdots & \boxed{A_0} \\ 0 & \end{pmatrix} S_1 = \begin{pmatrix} \lambda & ** \cdots ** \\ 0 & \\ \vdots & \boxed{S_0^{-1}A_0S_0} \\ 0 & \end{pmatrix} = \begin{pmatrix} \lambda & ** \cdots ** \\ 0 & \\ \vdots & \boxed{T_0} \\ 0 & \end{pmatrix}$$

is upper-triangular. Since this matrix is

$$S_1^{-1}(S^{-1}AS)S_1 = (SS_1)^{-1}A(SS_1),$$

the proof is complete. ■

11.3 Cramer's Rule

11.3.1 Solution of Regular Linear Systems

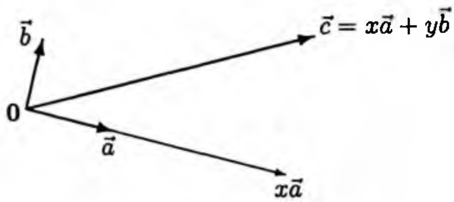
Here is a method for the resolution of regular linear systems of size $n \times n$, based on determinants. To emphasize its geometrical meaning, let us start by the 2×2 case

$$\begin{cases} a_1x + b_1y &= c_1 \\ a_2x + b_2y &= c_2. \end{cases}$$

This system can be written in vector form

$$x\vec{a} + y\vec{b} = \vec{c},$$

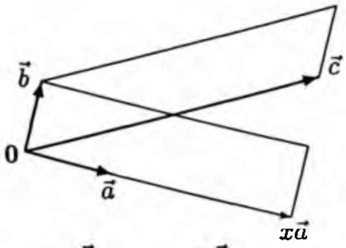
and is regular when $\vec{a}$ and $\vec{b}$ are linearly independent, which we assume. We have to find the components of the given vector $\vec{c}$ in the basis $\vec{a}$, $\vec{b}$.



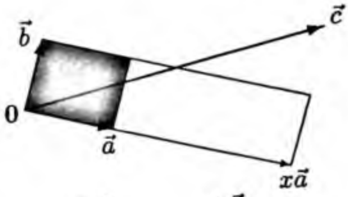
Using the 2-dimensional volume form (area) D, the following pictures illustrate the fact that the solution for x is a quotient of two determinants

$$x = \frac{D(\vec{c}, \vec{b})}{D(\vec{a}, \vec{b})},$$

hence as a quotient of areas of parallelograms.



$D(\vec{c},\vec{b}) = D(x\vec{a},\vec{b})$



$D(x\vec{a},\vec{b}) = xD(\vec{a},\vec{b})$

A regular 3×3 system

$$(S) \qquad \begin{cases} a_1x + b_1y + c_1z = d_1 \\ a_2x + b_2y + c_2z = d_2 \\ a_3x + b_3y + c_3z = d_3. \end{cases}$$

can be treated similarly using volumes instead of areas. In vector form, we have to determine the components of $\vec{\mathbf{d}}$ in the basis $\vec{\mathbf{a}}$, $\vec{\mathbf{b}}$, and $\vec{\mathbf{c}}$

$$x\vec{\mathbf{a}} + y\vec{\mathbf{b}} + z\vec{\mathbf{c}} = \vec{\mathbf{d}}.$$

In order to eliminate y and z, we can multiply (dot product) this equation by a vector which is orthogonal to both $\vec{\mathbf{b}}$ and $\vec{\mathbf{c}}$. Hence let us multiply by $\vec{\mathbf{b}} \wedge \vec{\mathbf{c}}$:

$$x\vec{\mathbf{a}} \cdot (\vec{\mathbf{b}} \wedge \vec{\mathbf{c}}) = \vec{\mathbf{d}} \cdot (\vec{\mathbf{b}} \wedge \vec{\mathbf{c}}).$$

We recognize scalar triple products

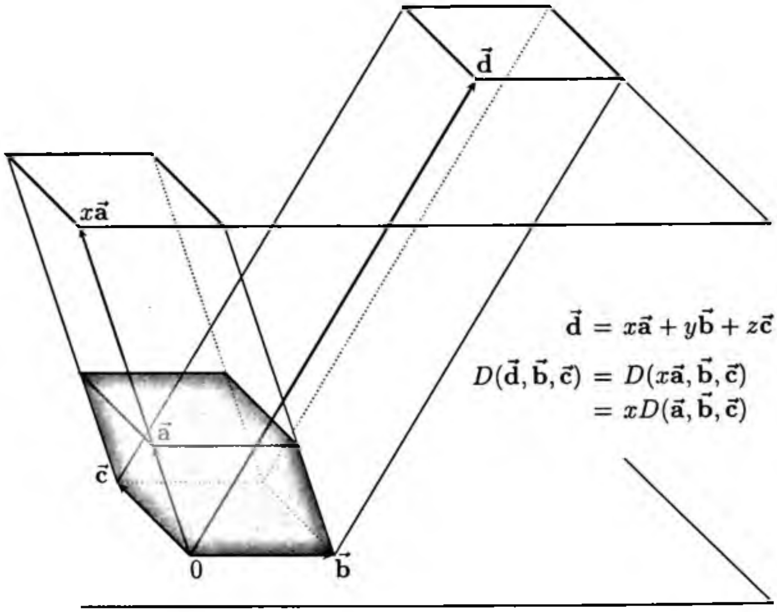
$$xD(\vec{\mathbf{a}}, \vec{\mathbf{b}}, \vec{\mathbf{c}}) = D(\vec{\mathbf{d}}, \vec{\mathbf{b}}, \vec{\mathbf{c}}),$$

and deduce

$$x = \frac{D(\vec{\mathbf{d}}, \vec{\mathbf{b}}, \vec{\mathbf{c}})}{D(\vec{\mathbf{a}}, \vec{\mathbf{b}}, \vec{\mathbf{c}})}$$

(the denominator does not vanish since $\vec{\mathbf{a}}, \vec{\mathbf{b}}, \vec{\mathbf{c}}$ are independent). Similar formulas are found for the variables y and z.

Just as the above picture gives the geometrical interpretation for the solution x of a 2×2 system (as a quotient of two areas), we can make a 3-dimensional illustration of the found formula for the solution x of a 3×3 system (as a quotient of two volumes).



Here is the generalization for regular $n \times n$ systems.

Theorem. Let A be an $n \times n$ invertible matrix, having columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$. Then the solution $\mathbf{x}$ of the linear system $A\mathbf{x} = \mathbf{y}$ has the components

$$x_j = \frac{f(\mathbf{a}_1, \ldots, \mathbf{y}, \ldots, \mathbf{a}_n)}{f(\mathbf{a}_1, \ldots, \mathbf{a}_j, \ldots, \mathbf{a}_n)} \quad (1 \leqslant j \leqslant n),$$

where $f \in \mathcal{D}(\mathbf{R}^n)$ is any volume form in $\mathbf{R}^n$.

PROOF. By block multiplication, we have

$$\mathbf{y} = A\mathbf{x} = (\mathbf{a}_1, \ldots, \mathbf{a}_n)\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \sum_{1 \leqslant k \leqslant n} x_k \mathbf{a}_k.$$

For any volume form f, we can compute $f(\mathbf{a}_1, \ldots, \mathbf{y}, \ldots, \mathbf{a}_n)$ (where $\mathbf{y}$ is placed in the jth position) as follows

$$\begin{aligned} f(\mathbf{a}_1, \ldots, \mathbf{y}, \ldots, \mathbf{a}_n) &= f(\mathbf{a}_1, \ldots, \Sigma_k x_k \mathbf{a}_k, \ldots, \mathbf{a}_n) \\ &= f(\mathbf{a}_1, \ldots, x_j \mathbf{a}_j, \ldots, \mathbf{a}_n) \\ &= x_j\, f(\mathbf{a}_1, \ldots, \mathbf{a}_j, \ldots, \mathbf{a}_n). \end{aligned}$$

The announced formula follows since $f \neq 0$ does not vanish on any basis. ∎

In particular, we may take the normalized volume form $f = D = D_n$, namely the n-dimensional determinant, and obtain the usual Cramer rule

$$x_j = \frac{\det B_j}{\det A},$$

where the matrix B_j is obtained from A by replacement of its jth column by $\mathbf{y}$.

11.3.2 Inversion of a Matrix

Let A be an invertible matrix, say of size $n \times n$. We are looking for the inverse of A. As usual, the columns of this inverse are the images $A^{-1}\mathbf{e}_j$ of the basis vectors. Let us determine the first column $\mathbf{x} = A^{-1}\mathbf{e}_1$ only (it is typical). It is the solution of the system $A\mathbf{x} = \mathbf{e}_1$ and we may use the Cramer rule for this purpose. Let B_j denote the matrices

$$B_j = \begin{pmatrix} a_{11} & \dots & 1_j & \dots & a_{1n} \\ a_{21} & \dots & 0_j & \dots & a_{2n} \\ \vdots & & & & \vdots \\ a_{n1} & \dots & 0_j & \dots & a_{nn} \end{pmatrix} \qquad (1 \leqslant j \leqslant n),$$

whose determinants give the numerators in the Cramer rule. We can expand the determinant of B_j according to its jth column (having only one nonzero entry)

$$\det B_j = (-1)^{j+1} \det A_{1j},$$

where A_{1j} denotes the sub-matrix of size $(n-1) \times (n-1)$ obtained by erasing the row and column of A containing a_{1j}. Hence the first column of A^{-1} is

$$a'_{j1} = (-1)^{1+j} \frac{\det A_{1j}}{\det A} \qquad (1 \leqslant j \leqslant n).$$

More generally, it is easy to see that the entries of A^{-1} are

$$a'_{ji} = (-1)^{i+j} \frac{\det A_{ij}}{\det A}.$$

Note the transposition between the entries of the inverse and the cofactors.

Examples. (1) If $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is invertible, then $\delta = ad - bc \neq 0$, and

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

(2) An $n \times n$ matrix having integral coefficients and determinant ± 1 has an inverse with integral coefficients too.

(3) The *Hilbert matrices*

$$H_n = \begin{pmatrix} 1 & 1/2 & \dots & 1/n \\ 1/2 & 1/3 & & \vdots \\ \vdots & & \ddots & \vdots \\ 1/n & \cdots & \cdots & 1/(2n-1) \end{pmatrix}$$

are invertible. It can be shown that $\det H_n$ is the inverse of an integer. For example

$$\det H_3 = 1/2160, \quad \det H_4 = 1/6048000, \quad \det H_5 = 1/266716800000.$$

More precisely, it is known that the Hilbert matrices have integral inverses. For example (a minicomputer TI92 is useful here!)

$$H_4^{-1} = \begin{pmatrix} 16 & -120 & 240 & -140 \\ -120 & 1200 & -2700 & 1680 \\ 240 & -2700 & 6480 & -4200 \\ -140 & 1680 & -4200 & 2800 \end{pmatrix},$$

$$H_5^{-1} = \begin{pmatrix} 25 & -300 & 1050 & -1400 & 630 \\ -300 & 4800 & -18900 & 26880 & -12600 \\ 1050 & -18900 & 79380 & -117600 & 56700 \\ -1400 & 26880 & -117600 & 179200 & -88200 \\ 630 & -12600 & 56700 & -88200 & 44100 \end{pmatrix}.$$

The determinant of H_n becomes very small when the size n increases: $\det H_9 \approx 10^{-42}$. On the other hand, the coefficients of the inverse get very large: H_9^{-1} has coefficients of the order of 10^{11}. This shows that the Cramer rule is not to be favored for the numerical computation of H_n^{-1}.

11.3.3 LU Factorizations: Necessary Condition

Let A be an invertible matrix, as in the previous subsection. Assume that it has a factorization $A = LU$ (Sec. 3.2.3) where

L : lower-triangular with 1's on the diagonal,
U : upper-triangular with pivots p_i on the diagonal.

Then, for each integer k between 1 and n ($k = n - 2$ in the picture below), multiplication by blocks (Sec. 3.3.1) with

$$\underbrace{\left(\begin{array}{cc|cc} \multicolumn{2}{c|}{A_k} & * & * \\ & & * & * \\ \hline * & * & * & * \\ * & * & * & * \end{array}\right)}_{A} = \underbrace{\left(\begin{array}{cc|cc} \multicolumn{2}{c|}{L_k} & 0 & 0 \\ & & 0 & 0 \\ \hline * & * & 1 & 0 \\ * & * & * & 1 \end{array}\right)}_{L} \underbrace{\left(\begin{array}{cc|cc} \multicolumn{2}{c|}{U_k} & * & * \\ & & * & * \\ \hline 0 & 0 & * & * \\ 0 & 0 & 0 & * \end{array}\right)}_{U}$$

shows that $A_k = L_k U_k$ is also an LU factorization (it has the same qualities). Hence

$$\det A_k = \underbrace{\det L_k}_{=1} \det U_k = \det U_k = \prod_{1 \leqslant i \leqslant k} p_i \neq 0$$

(we are assuming A invertible, so that there are n pivots $p_i \neq 0$). We even see here

$$p_k = \frac{\det A_k}{\det A_{k-1}} \quad (2 \leqslant k \leqslant n), \quad p_1 = a_{11}.$$

The sub-determinants $\det A_k$, corresponding to the square sub-matrices extracted from the top-left of A are the *principal minors* of A. Assume now conversely that all the principal minors of A are nonzero, and let us show that A has an LU factorization. The first principal minor is the coefficient a_{11}. If nonzero, it can be taken as first pivot p_1, and adding to the rows ρ_i $(i > 1)$ a multiple of the first one, we get a first equivalent

$$A \sim \begin{pmatrix} p_1 & * \cdots * \\ 0 & * \cdots * \\ \vdots & \vdots \\ 0 & * \cdots * \end{pmatrix}.$$

These special row operations keep all principal minors unchanged. In particular, if the second principal minor is nonzero, the second coefficient in the new second row can be taken as pivot p_2. Adding to the rows ρ_i $(i > 2)$ a multiple of the new second one, we get an equivalence

$$A \sim \begin{pmatrix} p_1 & * \cdots * \\ 0 & p_2 \cdots * \\ \vdots & \vdots \\ 0 & * \cdots * \end{pmatrix} \sim \begin{pmatrix} p_1 & * & * \cdots * \\ 0 & p_2 & * \cdots * \\ 0 & 0 & * \cdots * \\ \vdots & \vdots & \vdots \\ 0 & 0 & * \cdots * \end{pmatrix}.$$

Continuing in this way (use an induction on n), we see that without row exchange or scalar multiplication of rows, A is equivalent to an upper-triangular matrix U, hence $A = LU$ for some lower-triangular matrix L having 1's on its diagonal. This gives a satisfactory answer to the last question posed in Sec. 3.2.3.

11.4 Construction of Orthonormal Bases

Let $(\mathbf{v}_i)_{1 \leqslant i \leqslant n}$ be a finite family in an inner-product space E. Consider the matrix

$$G = (g_{ij}), \quad g_{ij} = (\mathbf{v}_i \mid \mathbf{v}_j).$$

The determinant of this symmetric matrix is the *Gram determinant* of the family

$$g = \mathrm{Gram}(\mathbf{v}_1, \ldots, \mathbf{v}_n) = \det(\mathbf{v}_i \mid \mathbf{v}_j).$$

This determinant detects the linear independence of the $(\mathbf{v}_i)$.

Proposition. *The following properties are equivalent:*

(*i*) *The family* $(\mathbf{v}_i)$ *is linearly independent*

(*ii*) $g = \det(g_{ij}) \neq 0$

(*iii*) $g = \det(g_{ij}) > 0$.

PROOF. We know (7.3.1) that there exists an orthonormal basis of the Euclidean subspace $V = \mathcal{L}(\mathbf{v}_1, \ldots, \mathbf{v}_n) \subset E$ generated by the $\mathbf{v}_i$. Take such a basis

$$\mathbf{e}_1, \ldots, \mathbf{e}_p \quad (p = \dim V \leqslant n),$$

and introduce the components of the $\mathbf{v}_j$ in this basis:

$$\mathbf{v}_j = \sum_{1 \leqslant k \leqslant p} a_{kj} \mathbf{e}_k \quad (1 \leqslant j \leqslant n).$$

This gives a matrix $A = (a_{ij})$ of size $p \times n$ and $G = (g_{ij}) = {}^t\!A\,A$:

$$\begin{aligned} g_{ij} &= (\mathbf{v}_i \mid \mathbf{v}_j) = \Big(\sum_{1 \leqslant \ell \leqslant p} a_{\ell i} \mathbf{e}_\ell \mid \sum_{1 \leqslant k \leqslant p} a_{kj} \mathbf{e}_k \Big) \\ &= \textstyle\sum_{\ell,k} a_{\ell i} a_{kj} (\mathbf{e}_\ell \mid \mathbf{e}_k) = \sum_{\ell,k} a_{\ell i} a_{kj} \delta_{\ell k} = \sum_k a_{ki} a_{kj}. \end{aligned}$$

The rank of G is less than or equal to p. If $p < n$, then $\det G = 0$. If $p = n$, the matrix A is a square matrix (having the same determinant as ${}^t\!A$), and

$$g = \det G = \det({}^t\!A\,A) = (\det A)^2 \geqslant 0.$$

Moreover $g = 0$ precisely when $\det A = 0$, namely when the vectors $\mathbf{v}_i$ are linearly dependent. ∎

The case of two vectors is already interesting. Indeed, in this case

$$g = (\mathbf{x} \mid \mathbf{x})(\mathbf{y} \mid \mathbf{y}) - (\mathbf{x} \mid \mathbf{y})^2 = \begin{vmatrix} (\mathbf{x} \mid \mathbf{x}) & (\mathbf{x} \mid \mathbf{y}) \\ (\mathbf{y} \mid \mathbf{x}) & (\mathbf{y} \mid \mathbf{y}) \end{vmatrix} \geqslant 0$$

is the Cauchy–Schwarz inequality once more! Observe that the equality can hold only if $\mathbf{x}$ and $\mathbf{y}$ linearly dependent (proportional).

Practically, the construction of orthogonal systems in an inner-product space can be done by an inductive procedure called *Gram–Schmidt orthogonalization*, that we explain now.

Theorem **(*Gram–Schmidt*).** *Let* $(\mathbf{v}_n)_{n \geqslant 0}$ *be a (finite or infinite) system of independent vectors in an inner-product space* E. *Then there is an orthogonal system* $(\mathbf{e}_n)$ *having the following property: For each* m

$$\mathcal{L}(\mathbf{e}_1, \ldots, \mathbf{e}_m) = \mathcal{L}(\mathbf{v}_1, \ldots, \mathbf{v}_m) \subset E.$$

Moreover, the vectors $\mathbf{e}_n$ *are unique up to multiplication by scalars.*

PROOF. We begin with $\mathbf{e}_0 = \mathbf{v}_0$ and proceed by indeterminate coefficients for the next one: $\mathbf{e}_1 = \mathbf{v}_0 + a_1\mathbf{v}_1 = \mathbf{e}_0 + a_1\mathbf{v}_1$ has to be orthogonal to $\mathbf{e}_0 = \mathbf{v}_0$. More generally, since the subspaces generated by the $(\mathbf{v}_i)_{i\leqslant m}$ and the $(\mathbf{e}_i)_{i\leqslant m}$ are the same, we have to take

$$\mathbf{e}_m = \mathbf{e}_0 + x_1\mathbf{e}_1 + \cdots + x_{m-1}\mathbf{e}_{m-1} + x_m\mathbf{v}_m,$$

with the orthogonality conditions

$$(\mathbf{e}_m \mid \mathbf{e}_j) = 0 \quad (j < m).$$

These conditions furnish a linear system for the unknown coefficients. The problem is to solve this system at each step. Thanks to the theory of determinants, there is an elegant way to give the result. Consider the symbolic determinant (10.4.4)

$$\begin{vmatrix} (\mathbf{v}_0 \mid \mathbf{v}_0) & \cdots & (\mathbf{v}_0 \mid \mathbf{v}_m) \\ \vdots & & \vdots \\ (\mathbf{v}_{m-1} \mid \mathbf{v}_0) & \cdots & (\mathbf{v}_{m-1} \mid \mathbf{v}_m) \\ \mathbf{v}_0 & \cdots & \mathbf{v}_m \end{vmatrix},$$

which represents a vector $\mathbf{e}_m \in E$. Expanding it according to its last row, we find

$$\mathbf{e}_m = g_m \cdot \mathbf{v}_m + \text{ a linear combination of the } \mathbf{v}_i \text{ for } i < m,$$

where g_m stands for the Gram determinant of the first m vectors $(\mathbf{v}_i)$. This determinant is nonzero by the preceding proposition (we are assuming that $\mathbf{v}_0, \ldots, \mathbf{v}_{m-1}$ are independent). We also see that

$$\mathcal{L}(\mathbf{v}_0, \ldots, \mathbf{v}_{m-1}, \mathbf{v}_m) = \mathcal{L}(\mathbf{v}_0, \ldots, \mathbf{v}_{m-1}, \mathbf{e}_m).$$

An easy induction shows that

$$\mathcal{L}(\mathbf{v}_0, \ldots, \mathbf{v}_{m-1}, \mathbf{v}_m) = \mathcal{L}(\mathbf{e}_0, \ldots, \mathbf{e}_{m-1}, \mathbf{e}_m).$$

Furthermore, distributing the inner products in the last row, we see that for $i < m$, $(\mathbf{e}_i \mid \mathbf{e}_m)$ is given by a determinant having two equal rows, hence zero: $\mathbf{e}_m$ is orthogonal to the previously constructed $\mathbf{e}_i$'s. ■

11.5 A Selection of Important Results

11.5.1 The Frobenius and Cayley–Hamilton Theorems

Any monic polynomial is (up to sign) the characteristic polynomial of a square matrix.

***Theorem** (**Frobenius**). Let $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$ be any monic polynomial of degree n. Then the so-called* companion matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & 1 \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1}, \end{pmatrix}$$

has characteristic polynomial $(-1)^n p(x)$.

PROOF. Indeed, let us compute

$$p_A(x) = \begin{vmatrix} -x & 1 & 0 & \cdots & 0 \\ 0 & -x & 1 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \vdots \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-1}-x \end{vmatrix}$$

according to an expansion with respect to its last row. We find

$$p_A(x) = (-a_{n-1} - x)(-x)^{n-1} + \sum_{0 \leqslant j \leqslant n-2} (-a_i)(-1)^{n+j+1} \det A_{jn},$$

where

$$\det A_{jn} = \begin{vmatrix} -x & 1 & \cdots & \cdots & 0 & 0 \\ 0 & -x & & \cdots & & \vdots \\ \vdots & & \ddots & 1 & & \vdots \\ \vdots & & & -x & 0 & 0 \\ 0 & 0 & \cdots & \cdots & 1_{n-j-1} & 0 \\ 0 & 0 & \cdots & 0 & * & \ddots \end{vmatrix} = (-x)^j.$$

Thus

$$p_A(x) = (-1)^n\Big(x^n + a_{n-1}x^{n-1} + \sum_{0 \leqslant j \leqslant n-2} a_j x^j\Big) = (-1)^n p(x),$$

and the theorem is proved. ■

***Theorem** (**Cayley–Hamilton**). Let A be a square matrix of size $n \times n$ and characteristic polynomial*

$$p_A(x) = \det(A - xI) = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1} + a_n x^n.$$

Then

$$p_A(A) = a_0 I + a_1 A + \cdots + a_{n-1}A^{n-1} + a_n A^n = O_n.$$

PROOF. Let $B = A - xI$, so that

$$\det B = p_A(x) = a_0 + a_1 x + \cdots + a_{n-1}x^{n-1} + a_n x^n$$

($a_n = (-1)^n$, $a_{n-1} = (-1)^{n-1}\operatorname{tr} A$, but we shall not use it). The inverse of B is

$$B^{-1} = \frac{1}{\det B} B_{\mathrm{cof}}$$

where B_{cof} has entries $b'_{ij} = (-1)^{i+j}\det B_{ji}$. This furnishes the identity

$$BB_{\mathrm{cof}} = B_{\mathrm{cof}}\, B = (\det B)\, I = p_A(x)\, I.$$

Since each entry of B_{cof} is $\pm\det B_{ji}$, where B_{ij} is obtained from $A - xI$ by deleting one row and one column, these entries are polynomials of degree at most $n-1$ in x, and we may write

$$B_{\mathrm{cof}} = B_0 + B_1 x + B_2 x^2 + \cdots + B_{n-1}x^{n-1}.$$

Now, the identity $p_A(x)\, I = B_{\mathrm{cof}}\, B$ furnishes explicitly

$$a_0 I + a_1 x I + a_2 x^2 I + \cdots + a_n x^n I = (B_0 + B_1 x + B_2 x^2 + \cdots + B_{n-1}x^{n-1})(A - xI).$$

In this *polynomial identity* (having matrix coefficients), we may substitute $x = A$, whence the result. More explicitly, we may identify the coefficients of like powers of x in the preceding polynomial identity, obtaining

$$\begin{aligned}
a_0 I &= B_0 A \\
a_1 I &= B_1 A - B_0 \\
a_2 I &= B_2 A - B_1 \\
&\ \ \vdots \\
a_{n-1} I &= B_{n-1} A - B_{n-2} \\
a_n I &= \phantom{B_{n-1}A} - B_{n-1}.
\end{aligned}$$

Let us multiply the second equation by A, the third equation by A^2, and so on (the last equation by A^n). Adding, we obtain a telescopic sum

$$p_A(A)I = B_0 A + (B_1 A^2 - B_0 A) + \cdots + (B_{n-1}A^n - B_{n-2}A^{n-1}) - B_{n-1}A^n,$$

namely $p_A(A) = O_n$, whence the theorem. ∎

This theorem shows that the nth power of the matrix A is a linear combination of the preceding powers of A. Multiplying by A, the $(n+1)$th power of A is a linear combination of the preceding ones. Hence an inductive method for computing all powers of A as linear combinations of $I, A, \ldots, A^{n-1}$. When A is invertible, we can also multiply the relation $a_0 I + a_1 A + \cdots + a_{n-1}A^{n-1} + a_n A^n = O_n$ by A^{-1}, obtaining

$$a_0 A^{-1} + a_1 I + \cdots + a_{n-1}A^{n-2} + a_n A^{n-1} = O_n.$$

Since $a_0 = \det A \neq 0$ in this case, we obtain an expression of the inverse as a linear combination of nonnegative powers of A:

$$(-A)^{n-1} + \operatorname{tr} A\,(-A)^{n-2} + \cdots = \det A \cdot A^{-1}.$$

This leads to

$$A^{-1} = \frac{1}{\det A}((-A)^{n-1} + \operatorname{tr} A\,(-A)^{n-2} + \cdots).$$

11.5.2 Restricting Scalars from C to R

We may view the complex field $\mathbf{C}$ as a real vector space, with basis 1 and i. This choice of basis leads to an isomorphism $\mathbf{R}^2 \xrightarrow{\sim} \mathbf{C}$ between these real vector spaces. Any $\mathbf{C}$-linear map $\mathbf{C} \to \mathbf{C}$ is given by multiplication $z \mapsto cz$ by some complex scalar $c \in \mathbf{C}$: Its determinant is c. This map is *a fortiori* $\mathbf{R}$-linear, and if $c = a + ib$ $(a, b \in \mathbf{R})$, its 2×2 matrix in the basis 1 and i, has for columns the components of

$$c\,1 = a + ib, \quad \text{resp. } c\,i = ai - b = -b + ai.$$

Hence this real matrix is

$$\begin{pmatrix} a & -b \\ b & a \end{pmatrix}.$$

The determinant of this $\mathbf{R}$-linear map is

$$\begin{vmatrix} a & -b \\ b & a \end{vmatrix} = a^2 + b^2 = c\bar{c} = |c|^2 \in \mathbf{R}.$$

This is the relation between the (complex) determinant of the $\mathbf{C}$-linear map

$$\begin{aligned} f_c : \mathbf{C} &\longrightarrow \mathbf{C} \\ z &\longmapsto cz, \end{aligned}$$

and the (real) determinant of the same map, but considered as an operator in the 2-dimensional real vector space $\mathbf{C}$. Here is a generalization.

Proposition. *Let $M \in \mathbf{M}_n(\mathbf{C})$ be an $n \times n$ matrix with complex entries and $f : \mathbf{C}^n \to \mathbf{C}^n$ the $\mathbf{C}$-linear map that it defines, with $\det M = \det_C f \in \mathbf{C}$. If f_R denotes the same map as f, but considered as $\mathbf{R}$-linear operator in $\mathbf{C}^n \cong \mathbf{R}^{2n}$, then*

$$\det f_R = |\det M|^2 \in \mathbf{R}.$$

PROOF. Suppose that

$$M = A + iB = (a_{jk} + ib_{jk}) \quad (a_{jk}, b_{jk} \in \mathbf{R})$$

(matrix of f in the canonical basis $(\mathbf{e}_j)_{1 \leqslant j \leqslant n}$ of $\mathbf{C}^n$). Take the real basis

$$\mathbf{e}_1, \ldots, \mathbf{e}_n, \mathbf{e}_{n+1} = i\mathbf{e}_1, \ldots, \mathbf{e}_{2n} = i\mathbf{e}_n$$

of $\mathbf{C}^n$. The matrix of the real linear map f in this basis has for first columns the components of

$$f(\mathbf{e}_k) = \sum_{1\leqslant j\leqslant n} (a_{jk} + ib_{jk})\mathbf{e}_j = \sum_{1\leqslant j\leqslant n} a_{jk}\mathbf{e}_j + \sum_{n<j\leqslant 2n} b_{jk}\mathbf{e}_j \quad (1 \leqslant k \leqslant n).$$

We infer that the real matrix attached to f has the block form

$$\begin{pmatrix} A & * \\ B & * \end{pmatrix} \in \mathbf{M}_{2n}(\mathbf{R}).$$

The last columns are formed by the components of

$$f(\mathbf{e}_{n+k}) = f(i\mathbf{e}_k) = if(\mathbf{e}_k)$$

(since we assume that f is $\mathbf{C}$-linear), hence these last columns are formed by the components of

$$f(\mathbf{e}_{n+k}) = i \sum_{1\leqslant j\leqslant n} (a_{jk} + ib_{jk})\mathbf{e}_j = \sum_{1\leqslant j\leqslant n} (-b_{jk} + ia_{jk})\mathbf{e}_j.$$

The real matrix attached to f has the block form

$$\begin{pmatrix} A & -B \\ B & A \end{pmatrix} \in \mathbf{M}_{2n}(\mathbf{R}).$$

(Notice that when M is real, $B = 0$ and all computations that follow are much simpler.) Consider the identity

$$\begin{aligned}\begin{pmatrix} A & -B \\ B & A \end{pmatrix}\begin{pmatrix} I_n & I_n \\ -iI_n & iI_n \end{pmatrix} &= \begin{pmatrix} A+iB & A-iB \\ B-iA & B+iA \end{pmatrix} \\ &= \begin{pmatrix} I_n & I_n \\ -iI_n & iI_n \end{pmatrix}\begin{pmatrix} A+iB & O \\ O & A-iB \end{pmatrix}.\end{aligned}$$

Taking determinants, we obtain

$$\begin{aligned}\begin{vmatrix} A & -B \\ B & A \end{vmatrix}\begin{vmatrix} I_n & I_n \\ -iI_n & iI_n \end{vmatrix} &= \begin{vmatrix} A+iB & O \\ O & A-iB \end{vmatrix}\begin{vmatrix} I_n & I_n \\ -iI_n & iI_n \end{vmatrix} \\ &= \det(A+iB)\det(A-iB)\begin{vmatrix} I_n & I_n \\ -iI_n & iI_n \end{vmatrix}.\end{aligned}$$

But

$$\begin{pmatrix} I_n & I_n \\ -iI_n & iI_n \end{pmatrix} \sim \begin{pmatrix} I_n & I_n \\ O & 2iI_n \end{pmatrix},$$

$$\begin{vmatrix} I_n & I_n \\ -iI_n & iI_n \end{vmatrix} = \begin{vmatrix} I_n & I_n \\ O & 2iI_n \end{vmatrix} = (2i)^n,$$

and we may simplify the preceding equality by this nonzero scalar, obtaining

$$\begin{vmatrix} A & -B \\ B & A \end{vmatrix} = \det(A+iB)\det(A-iB) = |\det(A+iB)|^2.$$

∎

This is the announced result.

11.6 Appendix

11.6.1 Back to AB and BA

Let us show that when A and B are rectangular matrices, A of size $m \times n$ (resp. B of size $n \times m$), then AB and BA have the same characteristic polynomial up to a power of $-x$. Hence all nonzero eigenvalues of AB and BA have the same algebraic multiplicities (in Sec. 6.5.1, we proved that they have the same geometric multiplicities). Here, AB is a square matrix of size $m \times m$, while BA is a square matrix of size $n \times n$.

A picture will help visualizing the situation

$$A = \begin{pmatrix} \boxed{\quad A \quad} \end{pmatrix} \qquad (m \text{ rows, } n \text{ columns})$$

and

$$B = \begin{pmatrix} \boxed{B} \end{pmatrix} \qquad (n \text{ rows, } m \text{ columns}).$$

Let us consider the following product by blocks

$$\begin{pmatrix} AB & O \\ B & O \end{pmatrix} \underbrace{\begin{pmatrix} 1_m & A \\ O & 1_n \end{pmatrix}}_{=S} = \begin{pmatrix} AB & ABA \\ B & BA \end{pmatrix}$$

$$= \underbrace{\begin{pmatrix} 1_m & A \\ O & 1_n \end{pmatrix}}_{=S} \begin{pmatrix} O & O \\ B & BA \end{pmatrix}.$$

Since the matrix S is obviously invertible, we deduce that

$$\begin{pmatrix} AB & O \\ B & O \end{pmatrix} = S \begin{pmatrix} O & O \\ B & BA \end{pmatrix} S^{-1}.$$

These similar matrices have the same characteristic polynomial, namely

$$p_{AB}(x) \cdot (-x)^n = (-x)^m \cdot p_{BA}(x)$$

whence the affirmation.