

# 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

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

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) :

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

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

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

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

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

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

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

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

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

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

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

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:

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

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

$$\left( \begin{array}{ccccc|c} 1 & -1 & -1 & 0 & 0 & -2 \\ 0 & 1 & 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & -3/2 & -2 & 2 \\ 0 & 0 & 0 & -5 + 9/2 & -6 + 6 & -6 \end{array} \right).$$

We have now reached the system

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

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

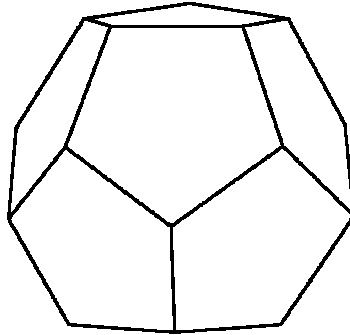
In any subdivision of the sphere consisting in hexagons and pentagons only, *the number of pentagons is fixed and equal to 12.*

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

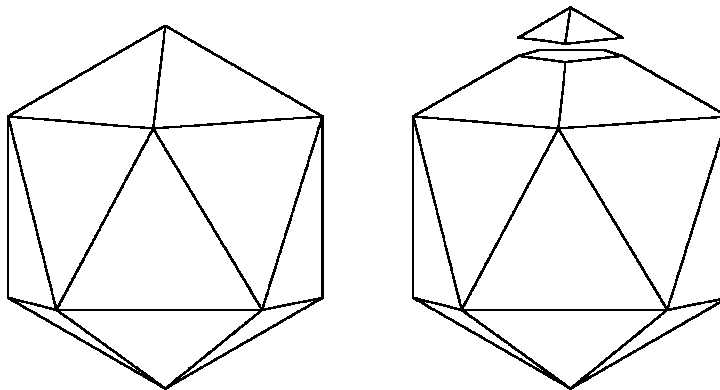


and project its twelve pentagonal faces on the surface of the sphere.

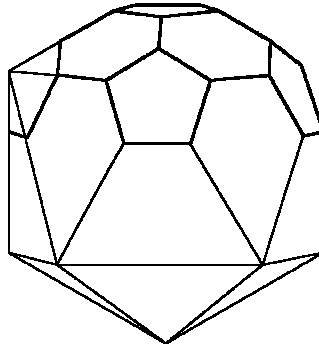
(b) Another solution

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

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

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

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.

(b1) 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

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

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

$$\left\{ \begin{array}{l} x = c + \frac{1}{2}A + \frac{1}{4}B + \frac{1}{4}C \\ y = c + \frac{1}{4}A - \frac{1}{4}C \\ z = c + \frac{1}{4}A - \frac{1}{4}B \\ u = c. \end{array} \right.$$

(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

$$\left\{ \begin{array}{l} x + y + z + u = A \\ x + y + z + u = B \\ x + y + z + u = C \\ x + y + z + u = D, \end{array} \right.$$

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

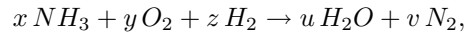
$$\left\{ \begin{array}{l} x + y + z - 3u = A \\ x + y - 3z + u = B \\ x - 3y + z + u = C \\ -3x + y + z + u = D, \end{array} \right.$$

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



where the proportions  $x, \dots, v$  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

$$\begin{cases} x & & & -2v & = 0 \\ 3x & & +2z & -2u & = 0 \\ & 2y & & -u & = 0. \end{cases}$$

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

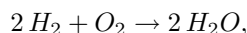
$$z = a - 3b.$$

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

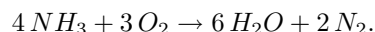
$$\begin{cases} x = 2b \\ y = \frac{1}{2}a \\ z = a - 3b \\ u = a \\ v = b \end{cases} \quad \text{also denoted} \quad \begin{pmatrix} x \\ y \\ z \\ u \\ v \end{pmatrix} = \begin{pmatrix} 2b \\ \frac{1}{2}a \\ a - 3b \\ a \\ b \end{pmatrix}.$$

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.<sup>1</sup> Two *basic solutions* appear. The first one corresponds to the choice  $u = 2$ ,  $v = 0$ , hence  $x = 0$  (no ammonia); it corresponds to the elementary reaction



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



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

<sup>1</sup>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<sup>12</sup>, 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

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

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

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

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  $x_1 = 0, x_2 = 0, \dots, x_n = 0$ .

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

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

where the coefficients  $p_1, p_2, \dots, p_r$  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*  $x_{j_1}, x_{j_2}, \dots, x_{j_r}$ . By definition, the rank  $r$  is less than or equal to  $m$  and  $n$

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

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  $r \leq m$ ). 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.<sup>2</sup> ■

<sup>2</sup>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  $x_{n-1} = 0, \dots$  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

$$1 \leq j_1 < j_2 < \dots < j_r \leq n.$$

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

$$\left( \begin{array}{cccccccc} 0 & \boxed{p_1} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\ 0 & & \boxed{p_2} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\ 0 & & & \boxed{p_3} & \text{---} & \text{---} & \text{---} & \text{---} \\ \vdots & & & & \ddots & & & \\ & & & & & \boxed{p_r} & \text{---} & \text{---} \\ 0 & \dots & & & & & & 0 \end{array} \right) \left. \vphantom{\begin{array}{c} \\ \\ \\ \\ \\ \\ \end{array}} \right\} \text{rank } r$$

ROW-REDUCED FORM

Here, the squares contain the pivot values  $p_i \neq 0$  ( $1 \leq i \leq r$ ), 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.

$$\left( \begin{array}{cccccccc} 0 & \boxed{1} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\ 0 & & \boxed{1} & \text{---} & \text{---} & \text{---} & \text{---} & \text{---} \\ 0 & & & \boxed{1} & \text{---} & \text{---} & \text{---} & \text{---} \\ \vdots & & & & \ddots & & & \\ & & & & & \boxed{1} & \text{---} & \text{---} \\ 0 & \dots & & & & & & 0 \end{array} \right) \left. \vphantom{\begin{array}{c} \\ \\ \\ \\ \\ \\ \end{array}} \right\} \text{rank } r$$

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  $\text{rref}(A)$ .

$$\left( \begin{array}{cccccccc} 0 & \boxed{1} & \text{---} & 0 & \text{---} & 0 & \text{---} & 0 & \text{---} \\ 0 & & \boxed{1} & \text{---} & 0 & \text{---} & 0 & \text{---} & 0 & \text{---} \\ 0 & & & \boxed{1} & \text{---} & 0 & \text{---} & 0 & \text{---} & 0 & \text{---} \\ \vdots & & & & \ddots & & & & & & \\ 0 & \dots & & & & \boxed{1} & \text{---} & 0 & \text{---} & 0 & \text{---} \end{array} \right) \left. \vphantom{\begin{array}{c} \\ \\ \\ \\ \\ \end{array}} \right\} \text{rank } r$$

#### REDUCED ROW-ECHELON FORM

This reduced row-echelon form corresponds to a system

$$\left\{ \begin{array}{l} 0 + x_{j_1} + \dots + \Sigma_1 = 0 \\ 0 + \dots + x_{j_2} + \dots + \Sigma_2 = 0 \\ 0 + \dots + \dots + x_{j_3} + \dots + \Sigma_3 = 0 \\ \vdots \\ 0 + \dots + \dots + x_{j_r} + \Sigma_r = 0, \end{array} \right.$$

where  $x_{j_1} = y_1$ ,  $x_{j_2} = y_2, \dots$ , and  $x_{j_r} = y_r$  are the pivot variables, while the sums  $\Sigma_j$  only involve the free variables  $y_{r+1}, \dots, y_n$ . (We use the capital Greek sigma  $\Sigma$  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

$$\left\{ \begin{array}{l} y_1 = x_{j_1} = -\Sigma_1 \\ y_2 = x_{j_2} = -\Sigma_2 \\ \vdots \\ y_r = x_{j_r} = -\Sigma_r. \end{array} \right.$$

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,  $\xi$  is the pivot variable.

### 1.3 Elimination Algorithm

An *algorithm*<sup>3</sup> 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:

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

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:  $d \geq 1$  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.

---

<sup>3</sup>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

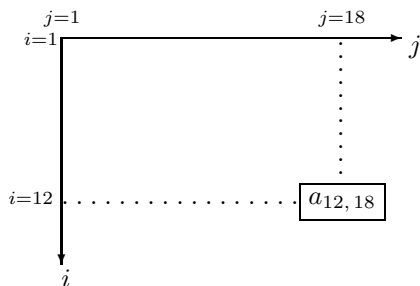
$$(S) \quad \begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ \vdots & \vdots & \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n = b_m. \end{cases}$$

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*

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

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

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

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

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

The system has a solution precisely when

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

or equivalently when

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

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

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

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

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

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

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

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

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

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.

**Variation.** 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

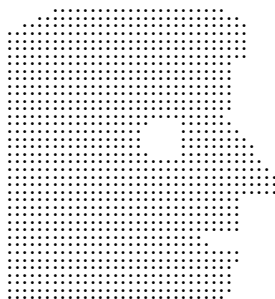
$$\begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n - b_1z = 0 \\ a_{21}x_1 + \cdots + a_{2n}x_n - b_2z = 0 \\ \vdots \qquad \qquad \qquad \vdots \qquad \qquad \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n - b_mz = 0. \end{cases}$$

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  $D_\varepsilon$  of  $D$ . Let us call *interior* vertices those having four neighbors (conveniently called North, East, South, and West) in  $D_\varepsilon$ , while the *boundary* vertices are those having less than four neighbors in  $D_\varepsilon$ . Here is an example of a discretization of a domain.



We are looking for a function  $f$  (potential) defined on the finite set  $D_\varepsilon$ , 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  $x_i = f(P_i)$  ( $1 \leq i \leq N$ ) 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

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

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  $P_i, \dots$  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  $x_i$  ( $1 \leq i \leq N$ ). *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  $N$ th 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

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

We find relations in the form

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

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  $x_1, x_2, \dots$ . The divisibility condition at the  $i$ th 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

$$x_0 = cN^N, \quad x_1 = \dots$$

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

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

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  $c = 1$

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

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

### 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 may 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  $\approx 10\text{\AA}$  ( $1\text{\AA} = 10^{-10}$  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),  $C_{240}, \dots$  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 (amoebae, 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

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

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

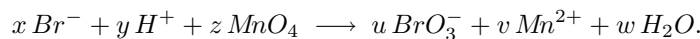
$$f(n) = 1 + \text{average of } (f(n-1), f(n+1)) \quad (1 \leq n < N).$$

(a) Check that the function  $h$  defined by  $h(n) = -n^2$  is a particular solution of the required functional equation.

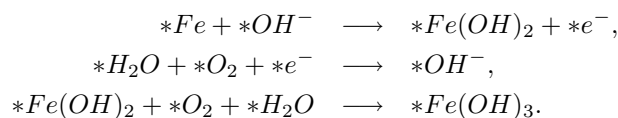
(b) All functions  $g$  of the form  $g(n) = An + B$  satisfy the associated homogeneous conditions.

(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



7. Same as before for the coupled reactions



8. Consider the following linear system in  $n$  variables

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

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

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

10. Let  $P_1, P_2, \dots, P_n$  be  $n$  given points in the space  $\mathbf{R}^3$ . Is it always possible to find disjoint balls  $B_i$  with center  $P_i$  ( $1 \leq i \leq n$ ) 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  $\alpha$ ,  $\beta$ , 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