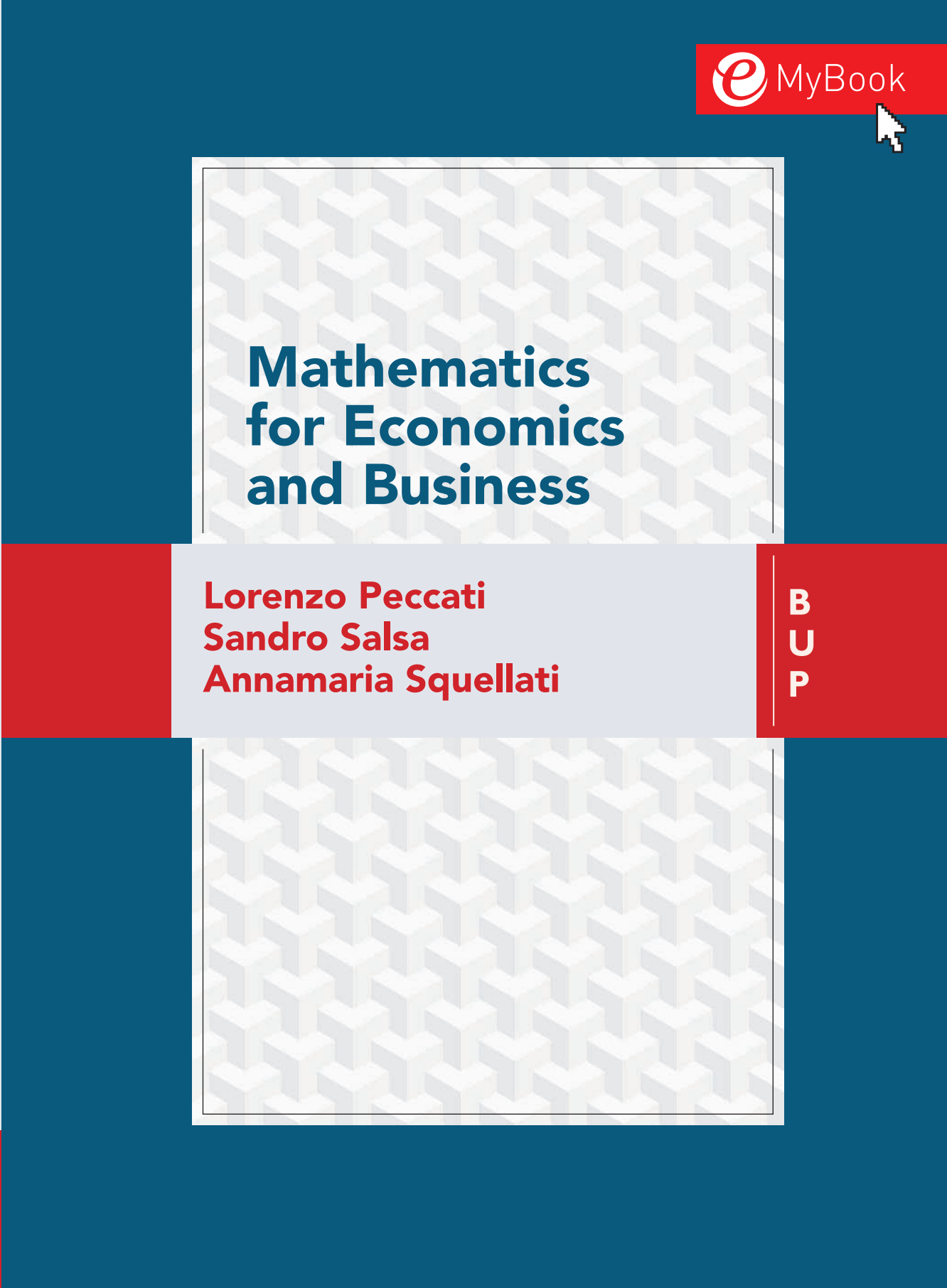

MyBook
Mathematics for Economics and Business
Lorenzo Peccati
Sandro Salsa
Annamaria Squellati
BUP

Mathematics for Economics and Business

Lorenzo Peccati
Sandro Salsa
Annamaria Squellati

BUP

Additional resources are available online via MyBook:
http://mybook.egeaonline.it

Translation: Guido Osimo (coordination), Sherren Hobson (revision), Gino Favero, Fabio Tonoli, Maria Beatrice Zavelani Rossi

EGEA S.p.A.
Via Salasco, 5 - 20136 Milan, Italy
Phone + 39 02 5836.5751 - Fax +39 02 5836.5753
egea.edizioni@unibocconi.it - www.egeaeditore.it

First Edition: September 2016

ISBN 978-88-99902-10-0

Print: Digital Print Service, Segrate (Milan)

Contents

Preface

The textbook *Mathematics for Economics and Business* is a translation of the Italian textbook *Matematica per l'Economia e l'Azienda* by the same Authors, and maintains the general outlines of its Italian counterpart.

We have selected some topics which we consider to be fundamental and mandatory for our students:
— the knowledge of Calculus, for functions of one and two variables;
— the use of Calculus in optimization;
— the notion of integral for functions of one variable;
— the language and the elementary techniques of Linear Algebra;
— the basics of Financial Calculus.
Several preliminary examples from applied sciences (mainly from Economics) introduce the theoretical aspects. We have tried to avoid an excessive formalism, in order to quickly reach the fundamental concepts.

The result is a textbook which is tailored for those educational programs which include a first (and perhaps only) course of Mathematics, in particular for those in Economics and Management.

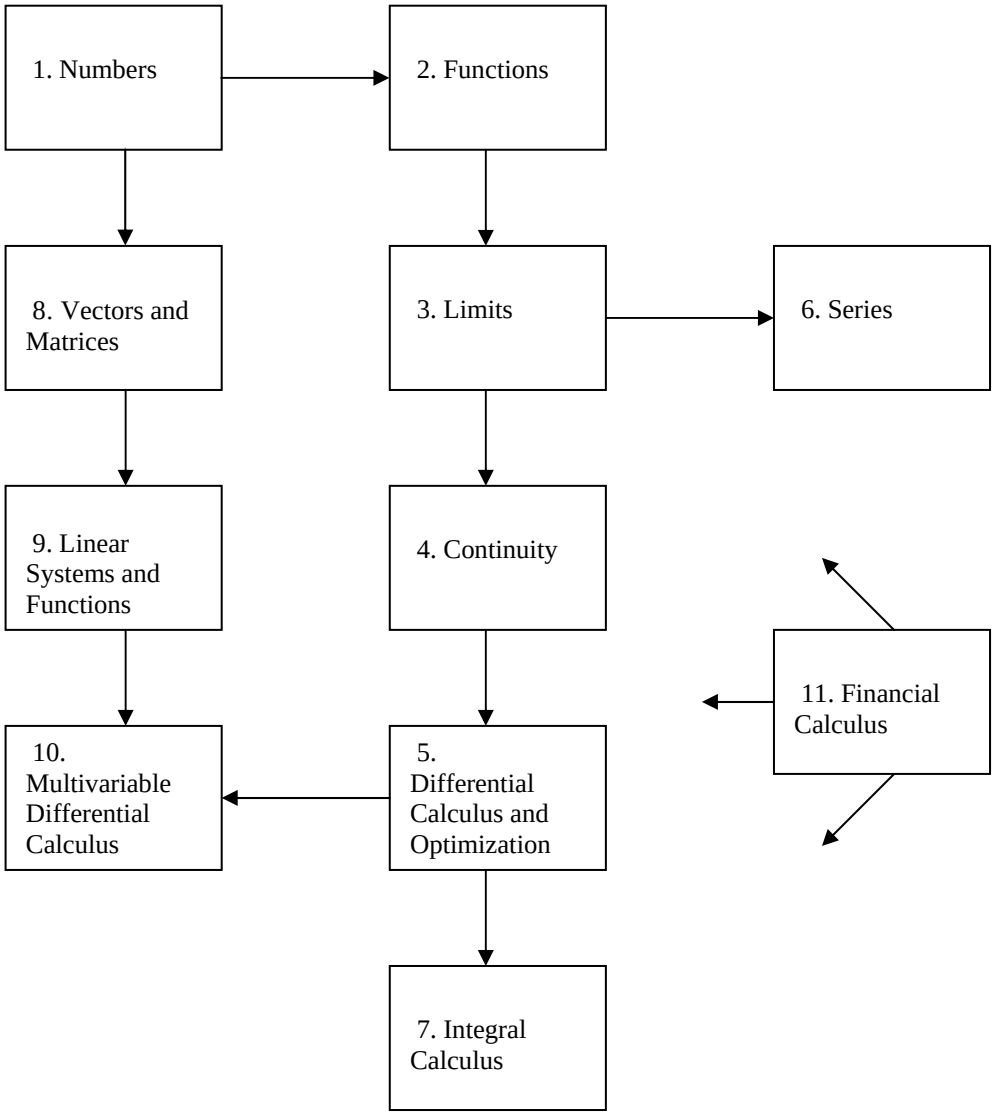
At the end of each Chapter a number of proposed exercises should be useful as a comprehension test. The solutions are available online via MyBook:

http://mybook.egeaonline.it

The Authors

Structure of the book

The following diagram shows the relations among Chapters and suggests a way to use this textbook. The first Chapter is an introduction to the whole book. Chapter 11 is a short introduction to Financial Calculus, and there are frequent links to it throughout the book.



1

Numbers

The aim of this chapter is to review certain mathematical concepts and tools which, in various ways, have been "travel mates" throughout the entire course of the reader's studies. The presentation follows a logical, inductive path which we can summarise as follows:

• A quick review of *numerical sets*, from *natural integers* to *relative integers* and from *rational* to *real* numbers. The aim is to emphasise their main properties and limitations, in order to avoid macroscopic errors. In particular, this is an appropriate point to consider:

— Roots, logarithms and exponentials, whose definitions are briefly analysed.

— The summation symbol, and the formulae for sums of arithmetic and geometric progressions.

• From numerical sets, the topic shifts to the general concept of *set* at an elementary level. The aim is to acquaint the reader with a language which has become indispensable. After considering various operations on sets, we cover:

— the Cartesian product, with the natural link with plane Cartesian coordinate geometry and some connected formulae;

— the concept of the *cardinality of a set*, together with the fundamental distinction between *finite* and *infinite* sets.

— the first elements of *combinatorial analysis*, with reference to finite sets;

— the concepts of *equipotent sets*, *countability* and *cardinality of the continuum*.

1.1 Natural and relative integers

1.1.1 Natural integers

Leopold Kronecker, a mathematician (1823-1891), used to say that "natural numbers are God's creation" and that, being already well established in our minds, they need not be defined. Adopting Kronecker's position, we have no problem stating that the *natural numbers* (or *natural integers*) are

$$0,\ 1,\ 2,\ 3,\ 4,\ 5, ..., 100, ..., 2008, ...$$

Natural numbers can be added and multiplied together. The operations of sum and product satisfy the commutative and associative properties, and are linked together by means of the distributive property. We know that, instead of

$$3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3,$$

it is more convenient to write 3^8 and that, generally speaking, the *power* m^n denotes the product of n factors, all equal to m. Raising to a power satisfies the properties

$$(m \cdot n)^r = m^r \cdot n^r \qquad m^n \cdot m^r = m^{n+r} \qquad (m^n)^r = m^{nr} \tag{1.1}$$

with the convention that $m^0 = 1$ if $m \neq 0$. The expression 0^0 has no meaning. Listing the natural numbers implies an ordering in their size: a number m is *greater* than another number n if it comes "later" in the listing, and we write $m > n$. If, on the contrary, m comes before n, it is referred to as *smaller* and we write $m < n$. Moreover, we write $m \geq n$ (or $m \leq n$) to denote that m is *greater than or equal to* (respectively, *less than or equal to*) n. Comparisons between natural numbers satisfy three important properties:

— *reflexive:* $m \geq m$;
— *antisymmetric:* if $m \geq n$ and $n \geq m$, then $m = n$;
— *transitive:* if $m \geq n$ and $n \geq p$, then $m \geq p$,

where m, n, p are any three natural numbers.

Sums and products "get on well together" with the ordering property. Indeed, if $m \leq n$, then

$$m + r \leq n + r \qquad \text{and} \qquad m \cdot r \leq n \cdot r. \tag{1.2}$$

1.1.2 Relative integers

When going on a winter holiday in the mountains and checking the outside temperature, we already need numbers "with a sign". With the only exception of zero, we can associate to every natural number m the symbol "+" (*plus*) or "−" (*minus*) and transform it into a *positive* number $+m$ or into a *negative* number $-m$, called the *opposite* of m. The numbers obtained this way are called the *relative integers*.

As well as with natural numbers, it is possible to perform the operations of sum and product with relative integers – which still satisfy the commutative, associative and distributive properties. The sum of a number with the opposite of another one

is called the *difference* between the two numbers, and instead of $m+(-n)$ we simply write $m-n$. The result of such an operation is still a relative integer, and thus there are no restrictions in subtracting relative integers. Concerning the product, one has to keep in mind the *sign rules*:

$$+\cdot+=+, \qquad -\cdot-=+, \qquad -\cdot+=+\cdot-=-.$$

Comparisons are regulated by a new list, which extends infinitely both on the left and on the right of zero, and can be visualised by putting the relative numbers on an oriented straight line. To this extent, it is enough to choose the positions of 0 and 1 on a straight line (usually 1 is set to the right of 0). In such a way, a unit of measure (the length of the segment between 0 and 1) and the *positive* orientation of the line (going from 0 to 1) are determined.

As in the case of natural integers, the numbers coming "before" in the listing are called *smaller* than those which follow. This way, every negative number is smaller than any positive number. Among negative numbers, note that $-3<-2$, $-2005<-1000$ and, generally speaking, given two natural numbers m and n, the inequality $m<n$ is equivalent to $-m>-n$. It is easy to check that reflexive, antisymmetric and transitive properties still hold.

Again, this ordering of the relative numbers "gets on well together" with the sum and product operations. Indeed, if m,n are relative numbers with $m<n$ and $r>0$, then the two formulae (1.2) still hold. On the other hand, if $r<0$, the direction of the second inequality *gets switched*. For instance, multiplying by -2 both sides of the inequality $5>3$, we get $-10<-6$.

1.2 Rational numbers

The problem of sharing up a pie or an inheritance cannot be solved by means of integer numbers. Thus, *rational numbers* come into play, i.e., numbers which can be represented as *ratios* (*fractions*) of relative integers, where care is taken not to put the number 0 as the denominator.

There are infinitely many *equivalent* fractions *representing* the same rational number. For instance, $1/7$ is equivalent to $2/14$, to $3/21$, and so on. Among them, it is particularly convenient to consider the fractions *reduced to its lowest terms*, where the numerator and the denominator are prime with respect to each other (i.e., they have no common factor). Usually the latter is the fraction used to represent a rational number, to the point that it is common to say, for instance, "the rational number $1/7$", thus identifying the number and the fraction. Fractions with denominator 1 (and the ones equivalent to them) correspond to the relative integers.

Sums and products of rational numbers are calculated by means of their representation as fractions, by following the rules that every reader surely knows and we are not going to recall here. Commutative, associative and distributive properties continue to hold true. Yet, rational numbers suffer fewer restrictions than the relative integers.

Given any fraction m/n different from 0, by taking into consideration the *reciprocal* fraction n/m we have

$$\frac{m}{n} \cdot \frac{n}{m} = 1.$$

This allows us to define division between rational numbers, the only reservation being not to use 0 as the divisor. Dividing m/n by p/q simply means multiplying m/n by the reciprocal q/p, thus still getting a rational number.

Comparison between two rational numbers is possible. It is of course enough to consider comparisons between positive numbers because, as well as with relative integers, $r < s$ is equivalent to $-r > -s$ and every negative number is smaller than every positive number. So, how can we decide which is the greater between 13/5 and 17/6? In general, it is better to reduce both numbers to the same denominator and compare the numerators. Thus, for positive numbers,

$$\frac{m}{n} \geq \frac{p}{q} \qquad \text{is equivalent to} \qquad mq \geq np.$$

Since $78 = 13 \times 6 < 5 \times 17 = 85$, we can conclude that $13/5 < 17/6$. The properties shown on page 2, as well as relations (1.2), still hold true.

The abundance of operations and related properties which hold for the rational numbers can be summarised by saying that the rational numbers, as a whole, constitute an *ordered field*. The noun "field" expresses the fact that the operations work correctly, and the adjective "ordered" expresses the consistency of those operations with respect to the relations $\leq$ and/or $\geq$.

• *Geometrical representation of rational numbers. Density.* Rational numbers can be geometrically represented by means of dots on the same oriented straight line we previously used for the relative integers. Let us call O (origin) the dot chosen to place 0 and A the dot corresponding to 1 (to the right of O). The positive rational number m/n is then placed to the right of O, at the dot P such that the segment OP has measure m/n with respect to the length of the segment OA (the unit of measure). We say that m/n is the *abscissa* of the dot P.

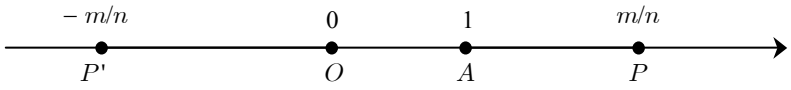


Figure 1.1. Rational numbers on the line

The number $-m/n$, the opposite of m/n, gets placed at P', the symmetric point of P with respect to the origin. In such a way, every rational number has a unique corresponding point on the straight line, which constitutes its geometrical image.

A profound difference from the relative integers is evident at a glance. Every integer sits on the line rigorously at a unit distance from the previous and the successive numbers. This is not the case for rational numbers. Indeed, between any

two rational numbers m/n and p/q, there always exists another rational number, for instance their arithmetic mean

$$\frac{1}{2}\left(\frac{m}{n}+\frac{p}{q}\right).$$

A moment's reflection is sufficient to be convinced that there are indeed infinitely many rational numbers between m/n and p/q! This property of being "thickly spread" on the line is expressed by saying that rational numbers are *dense* on the line. As we shall see soon, despite this property, they do not exhaust *all of* the points of a line, but still leave some (indeed, many) "holes".

• *Decimal representation.* When performing calculations with integer numbers, since the times of Leonardo Pisano, called *Fibonacci*[1], it is common to represent them using the so-called *positional* notation, mostly *in base* 10. This means that we choose the well-known symbols 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 (decimal digits) for the first ten integer numbers and, for instance, the written form 45123 is an abbreviation for $4\cdot 10^4+5\cdot 10^3+1\cdot 10^2+2\cdot 10+3\cdot 10^0$. It is indeed the positional notation which speeded up the arithmetic calculus and made it understandable by everybody. Try and think about multiplying 347 by 851 using the antique Romans' notation! Base ten is not the only one which gets used: for computers, the best bases are 2, 8, 16. The representational concept is just the same.

Rational numbers also have positional representations, in particular in base 10, and we can identify two types, illustrated in the two following examples. To represent 2/5 we just perform "2 divided by 5", thus getting

$$\frac{2}{5}=0.4.$$

By dividing 214 by 495, instead, we get

$$\frac{214}{495}=0.4323232\ldots$$

In the first case, a *finite* amount of numbers is required after the decimal point, whereas, in the second case, we see that the digits "32" repeat indefinitely. We say that the alignment is *recurrent* or *periodical*, that 32 is the *period* and we write $214/495=0.4\overline{32}$.

By dividing two integer numbers, it is not possible to get a recurrent decimal part with period 9. Nevertheless, nobody can deny $0.\overline{9}$ the *status* of number. Indeed, the decimal parts with period 9 can be considered as particularly bizarre ways to write a number: we can write $0.\overline{9}$ instead of 1, $12.3\overline{9}$ instead of 12.4, and so on.

1.3 Real numbers

We stated that rational numbers can be represented on the oriented straight line, but that they leave some "empty holes", meaning that there exist points on the

[1] "A son of Bonaccio" ($\simeq$1170-after 1240).

line which do not correspond to any rational number. The most immediate of these points is found by means of an argument which dates back at least to the times of Pythagoras (560-480 b.C.).

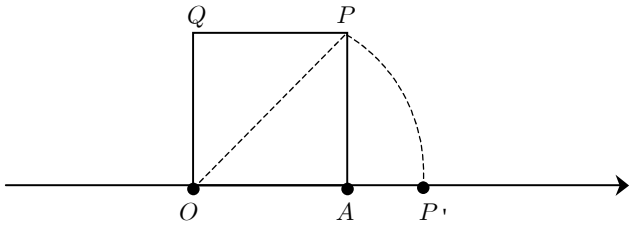


Figure 1.2. P' has no rational abscissa

Construct the square with unit edge and vertices O, A, P, Q as in Figure 2, draw the diagonal OP and transfer its length on the line with compasses, thus identifying the point P'. The latter *does not correspond to any rational number*. Indeed, the abscissa of P' coincides with the length of the diagonal OP, and the Pythagorean Theorem ensures that the square of such a length has to be equal to 2. Now, it is not possible to find a rational number whose square is 2. The process to prove this, though rather elementary, is quite interesting, and is a typical argument *by contradiction*: we suppose that such a number exists, and then deduce a contradiction.

Suppose that a given rational number, which we shall identify with its representing fraction m/n reduced to its lowest terms, had square 2: in a formula,

$$\frac{m^2}{n^2} = 2,$$

which can equally well be written as

$$m^2 = 2n^2. \tag{1.3}$$

The right hand side of (1.3) features an even number. So m^2, and thus m, has to be even as well. Now, since m is even, i.e., divisible by 2, m^2 turns out to be *divisible by* 4. On the other hand, since the fraction m/n is reduced to its lowest terms, n and thus n^2 must be odd numbers. As a consequence, $2n^2$ is divisible by 2 but not by 4, and thus it cannot be equal to m^2.

Pythagoreans had to conclude, with deep regret, that the diagonal of a square is *incommensurable* with its edge. Nevertheless, the diagonal was right there, as clearly visible as the edge, and there had to be some "number" measuring it somehow! Indeed there exists such a number, which we call the *square root of* 2 and denote by $\sqrt{2}$: it is an *irrational number*.

Irrational numbers find their places on the oriented straight line, in the "holes" left by the rational numbers. By considering the irrational and the rational numbers together, we get all of the *real* numbers. Clearly, every real number corresponds

to a unique point on the line and the converse is also true: we say that there is a *one-to-one correspondence* between real numbers and points on the line.

To get a slightly more satisfying definition of the real numbers, we can use decimal representations and define a *real number* by the representation

$$\pm p.a_1a_2a_3\ldots$$

with p is a natural integer and $a_1\ a_2\ a_3\ldots$ are decimal digits. Irrational numbers correspond to the representations which are neither finite nor recurrent. With an ordinary pocket calculator one can find the first digits in the decimal representation of the most common irrational numbers, such as:

$$\begin{aligned}\sqrt{2} &= 1.4142135624\ldots\\ \sqrt{3} &= 1.7320508076\ldots\\ \pi &= 3.1415926536\ldots\end{aligned}$$

It is possible to define sums, products and comparisons of real numbers, still preserving the same properties which hold for rational numbers. Thus, the real numbers constitute an *ordered field* as well. We restate concisely the properties of the operations *sum* (symbol $+$) and *product* (symbol $\cdot$). If a, b, c are any three real (or rational) numbers:

- *associative*

$$a+(b+c)=(a+b)+c \qquad a\cdot(b\cdot c)=(a\cdot b)\cdot c,$$

and thus one can simply write $a+b+c$ and $a\cdot b\cdot c$, discarding brackets;

- *commutative*

$$a+b=b+a \qquad a\cdot b=b\cdot a;$$

- there exist a *neutral element with respect to the sum* (*zero*) and a *neutral element with respect to the product* (*one*) such that

$$a+0=a \qquad a\cdot 1=a;$$

- for every element a there exists an *inverse element with respect to the sum* (called its *opposite* and denoted by $-a$) such that

$$a+(-a)=0;$$

- for every element $a\neq 0$ there exists an *inverse element with respect to the product* (called its *reciprocal* or *inverse* and denoted by a^{-1} or $1/a$) such that

$$a\cdot a^{-1}=1.$$

- The two operations are connected by the *distributive* property

$$(a+b)\cdot c=a\cdot c+b\cdot c$$

where, on the right hand side, it is understood that the product is "more binding" than the sum.

From the properties listed above follows, for instance, the possibility of solving *equations* through the well-known operations of "taking a term to the other side" (while changing its sign), or other similar ones, that the reader is surely familiar with. A *field* is then a good environment for solving equations.

The relation of *less than or equal to*, $\leq$, fulfills the *reflexive*, *antisymmetric* and *transitive* properties, and it is compatible with the arithmetic operations:

$$a \leq b \iff a + c \leq b + c \qquad \text{(compatibility w.r.t. the sum)}$$

and, if $c > 0$,

$$a \leq b \iff a \cdot c \leq b \cdot c \qquad \text{(compatibility w.r.t. the product).}$$

These two properties allow us to use together comparisons and arithmetic operations, thus permitting the solution of *inequalities* by means of the well-known rules of "taking a term to the other side", multiplications and so on, as mentioned above. An *ordered field* is then a good environment for solving inequalities.

The property distinguishing real numbers from rational ones, i.e. being in one to one correspondence with the points of a straight line, is called *completeness*. With this latter property, the *ordered field* of the real numbers becomes *complete*. The completeness of the field of the real numbers allows us to define operations which suffered too many restrictions in the field of the rational numbers, such as *root extraction*.

• *Absolute value or modulus. Distance between two points.* Let us evaluate the distance between two points on an oriented straight line. Let us start with the distance from the origin of a point P with abscissa x. Such a distance is of course the length of the segment OP, i.e., x if $x > 0$ and $-x$ if $x < 0$.

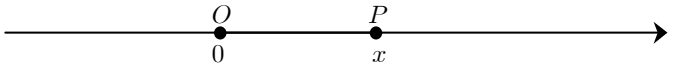


Figure 1.3.

The number defined by the formula

$$|x| = \begin{cases} x & \text{if } x \geq 0 \\ -x & \text{if } x < 0 \end{cases}$$

is called the *absolute value* (or the *modulus*) of x. The modulus of x thus represents the distance from the origin of the point with abscissa x.

Let us now consider the distance between two points P and Q with abscissae x and y respectively.

A simple check reveals that such a distance is $x - y$ if $x > y$, and $y - x$ if $x < y$. In brief:

$$\text{distance between } P \text{ and } Q = |x - y| \,.$$

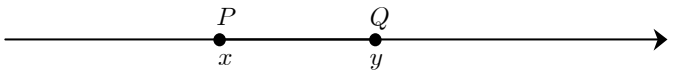


Figure 1.4.

An important inequality, which immediately follows from the definition of modulus of a real number, is worth pointing out.

Writing $|x| < a$, we mean that the distance of x from the origin is less than a, i.e., that

$$-a < x < a.$$

Since

$$-|x| \leq x \leq |x| \quad \text{and} \quad -|y| \leq y \leq |y|\,,$$

adding side by side we get

$$-(|x| + |y|) \leq x + y \leq |x| + |y|\,,$$

which amounts to the very important *triangle inequality*

$$\boxed{|x + y| \leq |x| + |y|\,.}$$

• *Arithmetic n-th roots.* As we know, Pythagoreans had to face the problem of finding a "number" which would equal 2 when squared. Since they did not know irrational numbers, they surrendered to the fact that the problem "had no solution". Since we know them, the problem of finding the possible values of x such that $x^2 = y$, with y a given real number, does not present any difficulty. If $y < 0$ there are no solutions (there is no real number whose square is negative). If $y = 0$ there is the only solution $x = 0$. If $y > 0$ there are two solutions: a positive one, call it x_1, and a negative one, $x_2 = -x_1$. The positive solution is called the *arithmetic square root of* y and denoted by $\sqrt{y}$ (the other one will be denoted by $-\sqrt{y}$).

Let us now look for the possible real numbers whose *cube* is y. The equation to be solved is $x^3 = y$, and we know that it has a unique solution x_1 for every y. It turns out that $x_1 = 0$ if $y = 0$, that $x_1 > 0$ if $y > 0$ and that $x_1 < 0$ if $y < 0$. For this solution, the symbol $\sqrt[3]{y}$ is used, called the *cubic root of* y (or also the *arithmetic cubic root*, in the case when $y > 0$).

What has been seen above for square and cubic roots can be extended without problems to the case of the equation $x^n = y$, treating separately the cases when n is odd and when n is even. Looking for solutions of such an equation corresponds to the operation of *extraction of the n-th root*. The fundamental point is stated in the following

Theorem 3.1. *Let y be a positive real number and $n \geq 1$ a natural integer. The equation*

$$x^n = y$$

admits one (and only one) positive solution, called the arithmetic n-th root of y and denoted by one of the symbols

$$\sqrt[n]{y} \qquad or \qquad y^{1/n}.$$

We add that when $n \geq 1$ is a natural integer we have $\sqrt[n]{0} = 0$, and that when y is a negative real number and $n \geq 1$ is an odd natural integer we can define the n-th root of y, $\sqrt[n]{y} = -\sqrt[n]{-y}$, which turns out to be negative.

Here is a question which often causes problems: among the following equalities, which is the correct one?[2]

$$\sqrt{x^2} = x, \quad \sqrt{x^2} = \pm x, \quad \sqrt{x^2} = |x|\,.$$

• *Powers with a real exponent.* Once the n-th roots have been defined, it is possible to define powers with a rational exponent by means of the following formulae. Let $a > 0$ and $r = m/n$, with m and n being two relatively prime natural integers. Set

$$a^r = \sqrt[n]{a^m} = \left(\sqrt[n]{a}\right)^m,$$

whereas, if $r < 0$, one sets

$$a^r = \frac{1}{a^{-r}}.$$

In some rare, yet well recognisable, cases it is also possible to allow for the base a to be less than or equal to zero.

Finally, it is possible to define powers with an irrational exponent, but *only when the base is positive*[3]. Powers with real exponents still satisfy properties (1.1). The following properties, which are useful for comparing powers with one another, also hold true:

$$\begin{aligned} &\text{for} \quad a > 1: \quad x_1 < x_2 \quad \text{if and only if} \quad a^{x_1} < a^{x_2} \\ &\text{for} \quad 0 < a < 1: \quad x_1 < x_2 \quad \text{if and only if} \quad a^{x_1} > a^{x_2}. \end{aligned}$$

• *Logarithms.* Consider the equality

$$2^5 = 32.$$

Given the base 2, in order to get 32 we have to use the exponent 5. We then say that 5 is the *logarithm in base* 2 *of* 32, and we write

$$5 = \log_2 32.$$

Thus, $\log_3(1/9) = -2$ (indeed, $3^{-2} = 1/9$), and $\log_{1/3} 9 = -2$ (as $(1/3)^{-2} = 9$).

To give a general definition of logarithm, consider the equation (with x unknown)

$$a^x = y. \tag{1.4}$$

[2] The third. The first is true only if $x \geq 0$, the second is meaningless.

[3] And, indeed, also in the case of positive exponent and null base: $0^\alpha = 0$ for every $\alpha > 0$.

Because of what we said about the powers with a real exponent, we are forced to choose $a > 0$. Then a^x always turns out to be *positive* and, as a consequence, if $y \leq 0$ there are no real solutions to equation (1.4). If furthermore $a = 1$, we have $a^x = 1^x = 1$ for every x, so that there are no solutions if $y \neq 1$, and all of the real numbers are solutions if $y = 1$.

Summarising, to make (1.4) a reasonable equation, we have to choose

$$\boxed{a > 0,\ \ a \neq 1 \ \text{ and } \ y > 0} \tag{1.5}$$

The following theorem then holds true.

Theorem 3.2. *Under conditions* (1.5)*, the equation* $a^x = y$ *admits a unique real solution called the* ***logarithm in base*** a ***of*** y *and denoted by the symbol*

$$\log_a y.$$

In other words, the logarithm in base a of y is defined by the important identity

$$\boxed{a^{\log_a y} = y}$$

The importance of logarithms from the point of view of calculations lies in some properties which we now recall. Numbers x and y are intended to be *positive*, and the base a, as it should be, is *positive* and $\neq 1$.

(a) $\log_a xy = \log_a x + \log_a y$

(b) $\log_a \dfrac{x}{y} = \log_a x - \log_a y$

(c) $\log_a \left(x^k\right) = k \log_a x$

(d) $\log_a x = \dfrac{\log_b x}{\log_b a} \qquad (b > 0,\ b \neq 1)$

(e) for $a > 1$: $\quad 0 < x_1 < x_2 \quad$ if and only if $\quad \log_a x_1 < \log_a x_2$,

for $0 < a < 1$: $\quad 0 < x_1 < x_2 \quad$ if and only if $\quad \log_a x_1 > \log_a x_2$.

As we shall have the chance to see from the next chapter on, the operations which we have introduced here will turn out to be important for the entire course.

1.4 Sum of terms in a progression

1.4.1 The summation symbol

Around 1820, the physicist and mathematician J. Fourier (1768-1830) introduced the symbol Σ (capital sigma), which is very convenient when dealing with complicated formulae. Suppose that we want to write the sum of the integer numbers from 1 to 26:

$$1 + 2 + 3 + \cdots + 25 + 26, \tag{1.6}$$

where the dots warn that the summation involves also the numbers from 4 to 24, not explicitly displayed. To denote (1.6), we can write

$$\sum_{n=1}^{26} n$$

which reads "the sum of n for n going from 1 to 26". Generally speaking, given a finite sequence of terms:

$$a_1,\ a_2, \ldots,\ a_n$$

(which reads "a sub one, a sub two, ... , a sub n", or simply "a-one, a-two, ..., a-n"), to denote their sum we can choose a letter, say s, as an index ranging from 1 to n, and write

$$\sum_{s=1}^{n} a_s$$

which reads "the sum of a (sub) s for s (going) from 1 to n". a_s is called the *general term*. The value of the sum *does not depend* on the name chosen for the index, but only on its *range*. For this reason, we say that the summation index is a *dummy* index.

Examples

4.1. To denote the sum

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots + \frac{1}{n^2}$$

we write

$$\sum_{s=1}^{n} \frac{1}{s^2} \quad \text{or} \quad \sum_{t=1}^{n} \frac{1}{t^2}.$$

4.2. The same sum can be written in many ways. For instance,

$$1 + 3 + 5 + \cdots + 29 + 31 \tag{1.7}$$

can be written as

$$\sum_{n=1}^{16} (2n - 1), \quad \text{or as} \quad \sum_{m=0}^{15} (2m + 1).$$

4.3. The sum

$$\sum_{m=0}^{100} 1$$

is a summation of one hundred and one terms (the index ranges from 0 to 100), *all equal to one*, and thus

$$\underbrace{1 + 1 + \cdots + 1}_{101 \text{ addends}} = 101.$$

4.4. The expression $(-1)^k$ is useful for dealing with alternating signs. Indeed, we have

$$(-1)^k = \begin{cases} +1 & \text{if } k \text{ is even} \\ -1 & \text{if } k \text{ is odd.} \end{cases}$$

As a consequence, the sum

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots - \frac{1}{16}$$

can be written as

$$\sum_{k=1}^{16} (-1)^{k+1} \frac{1}{k}.$$

The summation symbol is subject to some simple properties.

(a) $\sum_{k=1}^{n} c = cn$

(b) $\sum_{k=1}^{n} ca_k = c \sum_{k=1}^{n} a_k$ (distributive)

(c) $\sum_{k=1}^{n} (a_k + b_k) = \sum_{k=1}^{n} a_k + \sum_{k=1}^{n} b_k$ (commutative and associative)

(d) $\sum_{k=1}^{n} a_k = \sum_{k=1}^{m} a_k + \sum_{k=m+1}^{n} a_k$ $(m < n)$ (associative)

(e) $\sum_{k=1}^{n} a_k = \sum_{s=1+p}^{n+p} a_{s-p}$ (index change: $s = k + p$)[4].

Note that the sum $\sum_{k=n}^{m} a_k$ features $m - n + 1$ terms.

• *The product symbol.* Another symbol, analogous to $\sum$, can be used for the product: $\prod$ (capital pi). For instance, the product

$$1 \cdot 3 \cdot 5 \cdot 7 \cdot 9$$

can be written as

$$\prod_{s=1}^{5} (2s - 1)\,.$$

1.4.2 Sum of terms in an arithmetic progression

We say that n terms are in an *arithmetic progression* if the difference between any term and the preceding one is a constant, i.e., if such a difference is always equal to a given number, called the *common difference* of the progression. If we denote by a the first term and by c the common difference, the first n terms are

$$a, \quad a + c, \quad a + 2c, \ \ldots, \quad a + (n - 1)\, c.$$

[4] The last summation can also be written in the form

$$\sum_{k=1}^{n} a_k = \sum_{k=1+p}^{n+p} a_{k-p}$$

whence the rule: if one adds p to both the bounds (1 and n) and subtracts it from the general term index, the sum remains unchanged.

Their sum can be written with the summation symbol and calculated by means of the formula

$$\sum_{h=0}^{n-1}(a+hc)=\frac{n}{2}\left[2a+(n-1)\,c\right]. \tag{1.8}$$

The simplest example of an arithmetic progression is given by the positive integers themselves ($a=1$, $c=1$). We have

$$\sum_{k=1}^{n}k=1+2+\cdots+n=\frac{n\,(n+1)}{2}. \tag{1.9}$$

As another example,

$$3+7+11+\cdots+91=\sum_{k=0}^{22}(3+4k)=\frac{23}{2}\,(2\cdot 3+22\cdot 4)=1081.$$

1.4.3 Sum of terms in a geometric progression

The main feature of such a progression is that the ratio of any term with respect to the previous one is a constant, called the *common ratio* of the progression. We denote the common ratio by q. If the first term in the progression is 1, the second is obtained by $1\cdot q=q$, the third with another multiplication by q., i.e., $q\cdot q=q^2$, and so on. The sum of the first n terms, starting from 1, is then

$$1+q+q^2+q^3+\cdots+q^{n-1}, \tag{1.10}$$

which can be written in the concise form

$$\sum_{k=0}^{n-1}q^k.$$

For instance, if $q=1/2$, we get the sum

$$\sum_{k=0}^{n-1}\left(\frac{1}{2}\right)^k=1+\frac{1}{2}+\frac{1}{4}+\cdots+\frac{1}{2^{n-1}}.$$

If $q=1$, it is immediate that $\sum_{k=0}^{n-1}q^k=n$. If $q\neq 1$, the formula

$$\boxed{\sum_{k=0}^{n-1}q^k=\frac{q^n-1}{q-1}\qquad(q\neq 1)\,.} \tag{1.11}$$

holds true. To prove it, set

$$S=1+q+q^2+q^3+\cdots+q^{n-1}=\sum_{k=0}^{n-1}q^k \tag{1.12}$$

and multiply both sides by q, thus getting

$$qS = q + q^2 + q^3 + \cdots + q^n = \sum_{k=1}^{n} q^k. \tag{1.13}$$

Subtracting (1.12) from (1.13) side by side, we get

$$\begin{aligned}(q-1)S &= \sum_{k=1}^{n} q^k - \sum_{k=0}^{n-1} q^k = \\ &= q + q^2 + q^3 + \cdots + q^n - (1 + q + q^2 + q^3 + \cdots + q^{n-1}) = \\ &= q^n - 1,\end{aligned}$$

and, dividing by $q-1$, formula (1.11) is proved.

We can now easily prove that if the first term of the progression is a, the first n terms are:

$$a, \quad aq, \quad aq^1, \quad , \ldots , \quad , aq^{n-1},$$

and for the sum of the first n terms we get $\sum_{q=0}^{n-1} q^k = an$ when $q = 1$ and

$$\sum_{q=0}^{n-1} aq^k = a \cdot \frac{q^n - 1}{q - 1}$$

when $q \neq 1$.

Examples

4.5. One has

$$\sum_{k=0}^{n-1} \left(\frac{1}{2}\right)^k = \frac{\left(\frac{1}{2}\right)^n - 1}{\left(\frac{1}{2}\right) - 1} = 2\left[1 - \left(\frac{1}{2}\right)^n\right].$$

4.6. Evaluate

$$\sum_{k=2}^{n-1} (-3)^k.$$

It is the sum of a geometric progression with common ratio -3. In such a case, attention has to be paid to the starting value of the index, which in (1.11) is $k = 0$ but in this particular case is $k = 2$. No problem: by collecting $(-3)^2$ we can write

$$\sum_{k=2}^{n-1} (-3)^k = (-3)^2 \sum_{k=0}^{n-3} (-3)^k = 9\frac{(-3)^{n-2} - 1}{-3 - 1} = \frac{9}{4}\left[1 - (-3)^{n-2}\right].$$

Generally speaking, if $m < n$ and $q \neq 1$ we get

$$\sum_{k=m}^{n} q^k = q^m \sum_{k=0}^{n-m} q^k = q^m \frac{q^{n-m+1} - 1}{q - 1} = \frac{q^{n+1} - q^m}{q - 1}.$$

1.5 An outline of set theory

The meaning of the term *infinity* has always been a source of bitter quarrels among physicists, mathematicians and philosophers. The desire to solve *the infinity problem* in Mathematics was precisely one of the motivations which, towards the end of the nineteenth century, lead Georg Cantor (1845-1918) towards the foundation and the development of a theory which he himself called *Set theory*[5]. Its consequences have been, and still are, so deep and vast as to concern all of the sciences, both theoretical and applied. We are going to consider here the most superficial parts of Cantor's theory, just to introduce the language which is currently used in Mathematics.

What is a *set*? This is indeed one of the toughest questions (maybe *the* toughest one) in the whole theory. It is more or less like asking, when doing geometry in the plane, "what is a point"? A pragmatic attitude is to assume that all of us have in mind an idea about what a point is, and that such a concept needs no definition whatsoever. In any case, we already adopted a similar attitude when introducing the natural numbers. And, moreover, to give a correct answer we would need to resort to Euclid's axioms. Analogously, to define the concept of *set* we would need an axiomatic set-up which is much heavier than Euclid's framework. As a consequence, we introduce the concept of *set* by adopting an intermediate point of view which, after all, was the one originally adopted by Cantor himself.

1.5.1 Sets

A set is identified by explicitly declaring the *objects* (the *elements*) belonging to it, or a *property* which characterises them. The letters in the English alphabet

$$a,\ b,\ c,\ d, \ldots,\ w,\ x,\ y,\ z$$

constitute a set. Apart from the explicit itemising, the characterising property is "being a letter of the alphabet". The following are some examples of sets (of various types): the apple trees, the psychoanalysts, the irrational numbers, the time interval spent at the dentist's (this is a set of instants), the pixels in a monitor, the molecules of a given cat.

Sets are usually denoted by capital letters such as $A, B, X, Y \ldots$, while their elements are denoted by small letters $a, b, x, y \ldots$. When all of the elements of a set are explicitly listed, they are put between curly brackets. The set of all letters in the alphabet, then, is denoted by writing

$$\{a, b, c, d, \ldots, x, y, z\}\,.$$

The curly brackets, in this case, are meant to give an aggregation status to the elements. By writing

$$\{0, 1, 2\}$$

[5] Of course, he actually called it *Mengenlehre*.

we denote the set whose only elements are the integer numbers 0, 1 and 2. On the contrary, if we simply write 0, 1, 2 we intend the integer numbers 0, 1 and 2 each as an individual entity.

A typical symbol in Set theory, which is too important to be given up, is the symbol[6] $\in$, which denotes that an element *belongs* to a set. Writing $x \in \{0, 1, 2\}$ means that either $x = 0$ or $x = 1$ or $x = 2$. Of course, 2004 does not belong to $\{0, 1, 2\}$, and hence we write $2004 \notin \{0, 1, 2\}$.

Among the various sets there is a very special one, without any element, called the *empty set* and denoted by $\emptyset$.

1.5.2 Relations and operations with sets

Two sets are *equal* if they have the same elements. The sets

$$\{a, b, c\} \quad \text{and} \quad \{b, c, a\}$$

are equal: as we can see, *the ordering* with which the elements are listed *is not relevant*, but *only* the elements themselves matter.

Definition 5.1. *We say that A is a **subset** of B, or that A is **contained** (or **included**) in B, if every element in A is also an element of B, and we write*

$$A \subseteq B.$$

The inclusion symbol "$\subseteq$" reminds us of the symbol "$\leq$" of less than or equal to and, indeed, the fact that $A \subseteq B$ does not exclude the possibility that A and B coincide. If, on the other hand, we want to rule out this possibility, the symbol of *proper* (or *strict*) inclusion may be used, i.e.,

$$A \subset B,$$

which reads "A is strictly included in B". The set $\emptyset$ is strictly included in *every* other set. If $A \neq \emptyset$ and $A \subset B$, we also say that A is a *proper* subset of B. Every set A has as *improper* subsets A itself and $\emptyset$.

Definition 5.2. *The **union** of two sets A and B is the set, denoted by*

$$A \cup B,$$

made up of all of the elements belonging either to A or to B.

For short, introducing the symbol of *assignment* ":=", we can write

$$A \cup B := \{x : x \in A \text{ or } x \in B\},$$

which reads "A union B is defined as the set of the elements x such that[7] either x belongs to A **or** x belongs to B". The union of the two sets

$$A := \{\spadesuit, \heartsuit, \diamondsuit, \clubsuit\} \quad \text{and} \quad B := \{\nabla, \square, \clubsuit, \heartsuit\} \tag{1.14}$$

[6] Due to the Italian mathematician Giuseppe Peano (1858-1932).

[7] The colon inside the brackets then reads: "such that". Alternatively, the vertical bar | may be used.

is the set

$$\{\spadesuit,\ \diamondsuit,\ \nabla,\ \square,\ \clubsuit,\ \heartsuit\}\,.$$

The union of the set of the irrational numbers and the set of the rational numbers is the set of the real numbers.

Definition 5.3. *The **intersection** of two sets A and B is the set, denoted by*

$$A\ \cap B,$$

made up of all of the elements belonging both to A and to B.

As we did for the union, we can write for short

$$A\ \cap B := \{x : x \in A \text{ and } x \in B\}\,,$$

which reads "A intersection B is defined as the set of the elements x such that x belongs to A **and** x belongs to B". The intersection of the two sets in (1.14) is

$$\{\heartsuit,\ \clubsuit\}\,.$$

The intersection between the set of the even integer numbers and the set of the numbers divisible by 3 is the set of the numbers divisible by 6.

If two sets have an empty intersection, i.e., if $A \cap B = \emptyset$, they are called *disjoint*. The set of the rational numbers and the set of the irrational numbers are disjoint.

• *Power set.* It often happens that all of the sets under consideration are subsets of a common set U, called the *universe* set. The set of all (proper and improper) subsets of U is given the name of *power set* of U and is denoted by $\mathbb{P}(U)$ or 2^U.

To help your intuition, think about the toss of a die and consider as universe the set

$$U := \{1,\ 2,\ 3,\ 4,\ 5,\ 6\}$$

of all possible results. The power set of U is made up of 64 subsets: some of them are listed here below.

— $\emptyset$
— $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}$
— $\{1,2\}, \{1,3\}, \{1,4\}, \cdots, \{4,6\}, \{5,6\}$
— $\{1,2,3\}, \{1,2,4\}, \cdots, \{3,5,6\}\,\{4,5,6\}$
— $\{1,2,3,4\}, \{1,2,3,5\}, \cdots, \{2,4,5,6\}, \{3,4,5,6\}$
— $\{1,2,3,4,5\}, \{1,2,3,4,6\}, \cdots, \{1,3,4,5,6\}, \{2,3,4,5,6\}$
— $\{1,2,3,4,5,6\}\,.$

The family of these subsets represents the entirety of the events associated with the toss of a die. For instance, the subset $\{2,4,6\}$ is associated with the event "the result is an even number". Analogously, the event "the result is either an odd number or 2" corresponds to the set $\{1,2,3,5\}$, which is the union between $\{1,3,5\}$ and $\{2\}$. The empty set corresponds to an "impossible" event.

Definition 5.4. *For every subset A of U, the symbol A^c_U (or A^c or $\overline{A}$, when the universe set U is understood) denotes the **complement** set of A with respect to U, i.e., the set made up of all of the elements in U which do not belong to A.*

Thinking again about tossing a die, the complement set of $\{1, 2\}$ ("the result is either 1 or 2") is the set $\{3, 4, 5, 6\}$ ("the result is a number greater than 2").

Definition 5.5. *Given two sets A, B, the symbol $A \setminus B$ denotes the* ***difference*** *set between A and B, made up of the elements which belong to A but not to B.*

The difference between the sets in (1.14), which reads A *minus* B, is the set

$$A \setminus B = \{\spadesuit,\ \diamondsuit\}\,.$$

The difference between the set of the even integer numbers and the set of the integer numbers divisible by 3 is the set of all even integer numbers which are not multiples of 6.

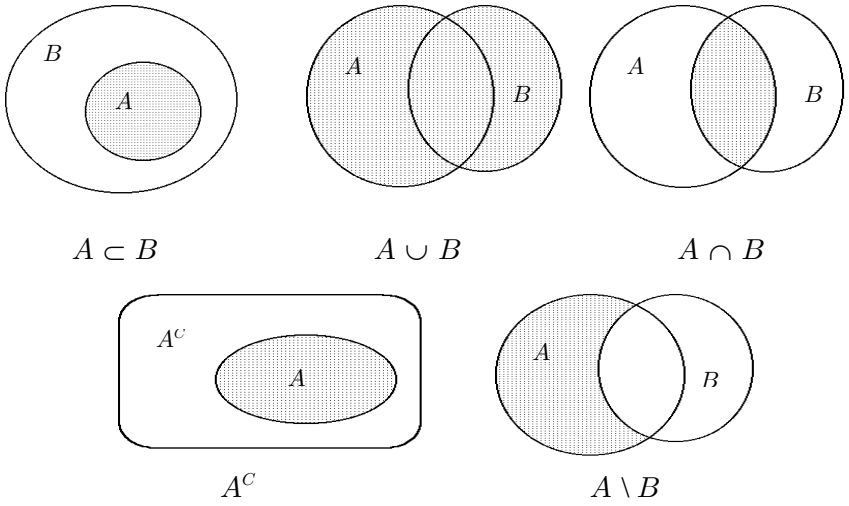


Figure 1.5. Venn diagrams

The meaning of inclusion, union, intersection, complement and difference can be represented by means of the so-called *Venn diagrams* (Figure 5). Note that the operations introduced above show strong analogies with the arithmetic operations, where the union plays the role of the sum, whereas the intersection plays that of the product. Indeed, the following *properties*, which are immediately verifiable and where A, B and C denote any three sets, hold true:

$$\begin{aligned} A \cup B &= B \cup A, \qquad A \cap B = B \cap A \qquad \text{(commutative)} \\ (A \cup B) \cup C &= A \cup (B \cup C)\,,\ (A \cap B) \cap C = A \cap (B \cap C) \qquad \text{(associative)} \end{aligned}$$

(as a consequence, it is possible to write simply $A \cup B \cup C$ and $A \cap B \cap C$),

$$\begin{aligned} (A \cup B) \cap C &= (A \cap C) \cup (B \cap C) \\ (A \cap B) \cup C &= (A \cup C) \cap (B \cup C) \end{aligned} \qquad \text{(distributive)}$$

The empty set behaves with respect to sets like the number zero does with respect to numbers:

$$A \cup \emptyset = A, \qquad A \cap \emptyset = \emptyset.$$

• *De Morgan's laws*[8]. These laws regulate the action of complementarity with respect to union and intersection.

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

In words: the complement set of the union is the intersection of the complement sets, the complement set of the intersection is the union of the complement sets. According to these laws, every property of the union of sets can be translated, switching to complement sets, into a property of the intersection.

1.5.3 Cartesian product

We have seen that the sets of two elements $\{a, b\}$ and $\{b, a\}$ coincide, because they have the same elements. In many practical situations we need to deal with *ordered* pairs, where the ordering in which the elements are written does matter. More precisely, given two sets A and B (not necessarily different), an *ordered pair* is a set, denoted by (a, b), obtained by choosing an element $a \in A$ and an element $b \in B$ in the specified order. Two ordered pairs (a, b), (a', b') are equal if

$$a = a' \quad \text{and} \quad b = b'.$$

Definition 5.6. *The set of all of the ordered pairs (a, b), with $a \in A$ and $b \in B$, is called the* ***cartesian product*** *of A and B and is denoted by $A \times B$.*

In formulae,

$$A \times B := \{(a, b) : a \in A \text{ and } b \in B\}.$$

Given the importance of the ordering in the pair (a, b), it should be clear that, whenever A is different from B, $A \times B \neq B \times A$. In the case when $A = B$, we write $A \times A = A^2$.

1.6 Sets of real numbers

To denote the sets of the integer, rational and real numbers, the following symbols are now widespread:

$\mathbb{N}$ natural integers,
$\mathbb{Z}$ relative integers,
$\mathbb{Q}$ rational numbers,
$\mathbb{R}$ real numbers.

[8] Augustus De Morgan (1806-1871), English mathematician.

Note that

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R}.$$

Some subsets of $\mathbb{R}$, which are called *intervals*, deserve special attention. Given two real numbers a, b with $a < b$, the set

$$[a, b] := \{x \in \mathbb{R} : \ a \leq x \leq b\}$$

is called the *closed and bounded interval with extremes a and b*. The specification "closed and bounded" comes from the fact that the extremes a and b belong to the set (whence *closed*) and at the same time provide a lower and an upper bound (whence *bounded*). Analogously, the set

$$(a, b) := \{x \in \mathbb{R} : \ a < x < b\}$$

is called an *open* (the extremes are not included) *and bounded interval*. The intervals

$$\begin{aligned} (a, b] &: = \{x \in \mathbb{R} : \ a < x \leq b\} \\ [a, b) &: = \{x \in \mathbb{R} : \ a \leq x < b\} \end{aligned}$$

are neither open nor closed. All of these intervals have a segment of a straight line as geometrical image. A half-line is instead the geometrical image of an *unbounded interval*, i.e., a set of the type

$$\begin{aligned} [a, +\infty) &: = \{x \in \mathbb{R} : \ x \geq a\} \qquad \text{(closed and unbounded to the right)}, \\ (a, +\infty) &: = \{x \in \mathbb{R} : \ x > a\} \qquad \text{(open and unbounded to the right)}, \end{aligned}$$

or

$$\begin{aligned} (-\infty, b] &: = \{x \in \mathbb{R} : \ x \leq b\} \qquad \text{(closed and unbounded to the left)}, \\ (-\infty, b) &: = \{x \in \mathbb{R} : \ x < b\} \qquad \text{(open and unbounded to the left)}. \end{aligned}$$

Consistently, we can write $\mathbb{R} = (-\infty, +\infty)$, which is the only unbounded interval whose geometrical image is the entire straight line. Particularly important intervals are those associated with the concept of a *neighbourhood of a point*.

Definition 6.1. *Given a point x_0 in $\mathbb{R}$, the open interval $(x_0 - a, x_0 + a)$, with $a > 0$, is called the* ***neighbourhood*** *of x_0 with radius a.*

A neighbourhood of a point x_0 is therefore an open and bounded interval centred at x_0, and its elements are all of the points whose distance from x_0 is less than a, i.e., all of the real numbers such that

$$|x - x_0| < a.$$

1.6.1 Maximum and minimum of a set

The possibility of comparing real numbers leads in a natural way to the concepts of *maximum* and *minimum* among the elements of a subset.

Definition 6.2. *A real number m is called the* ***maximum*** *(respectively, the* ***minimum****) of a set $A \subseteq \mathbb{R}$ if*

(i) $m \in A$ and

(ii) for every $a \in A$ one has $a \leq m$ (respectively, $a \geq m$).

Consider the interval $I = [-1, 3]$. The maximum of I is 3, and the minimum of I is -1. We shall use the following notation for the maximum and minimum of a set A:

$$\max A, \qquad \min A.$$

A maximum and a minimum do not always exist. It is enough to consider the intervals $I_1 = (-1, 3)$, $I_2 = [-1, 3)$, $I_3 = (-1, 3]$. I_1 has neither a maximum nor a minimum, as 3 and -1 satisfy condition (ii) but do not belong to I_1 and hence do not satisfy (i). Analogously, I_2 has no maximum and $\min I_2 = -1$, whereas I_3 has no minimum and $\max I_3 = 3$. The requirement that a maximum belong to the set is then very relevant.

We furthermore call a set $A \subseteq \mathbb{R}$ *bounded from above* (*bounded from below*) if it cannot extend indefinitely rightwards (respectively, leftwards) on the real line, i.e., *if it is possible to find a number h such that, for every $a \in A$, we have $h \geq a$ ($h \leq a$).* Finally, a set is called *bounded* if it is simultaneously bounded from above and from below.

1.7 The cartesian plane

The star among cartesian products is without any doubt $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$, the set of all ordered pairs of real numbers. Thanks to the one-to-one correspondence between real numbers and points on a straight line, it is possible to represent the elements of $\mathbb{R}^2$ as points on a plane, which takes the name of *cartesian plane*. In order to do so, we first fix two oriented straight lines, which are called the *cartesian axes* and usually (but not necessarily) are taken to be perpendicular to each other, as shown in Figure 6.

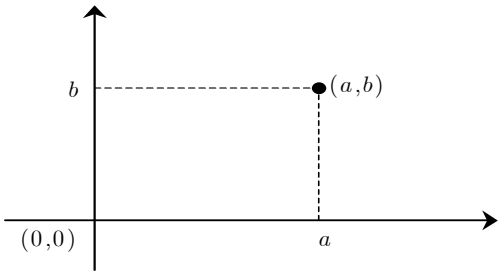


Figure 1.6. Coordinate system in the plane

The pair $(0, 0)$ corresponds to the intersection point between the axes (called the *origin*). On the horizontal axis, or *abscissae* axis, the first element of the ordered pair

is represented, whereas the second is placed on the vertical axis, or *ordinates* axis. We then say that *an orthogonal cartesian coordinate system*, or an *orthogonal reference system* has been set up in the plane. This way, a one-to-one correspondence is set up between points in the plane and ordered pairs of real numbers (the *coordinates* of the points), as shown in Figure 6.

The capability of "labeling" every point in the plane (a geometrical entity) with an algebraic object (the pair of its coordinates) has an enormous significance, both conceptual and practical. It indeed allows us to solve geometrical problems by means of algebraic methods, and conversely to solve algebraic problems by means of geometrical methods. This approach takes the name of *coordinate geometry*, a discipline where *every* geometrical entity, once the reference system is fixed, corresponds in a one-to-one way to an algebraic object.

• *Length of a segment.* Let us consider the segment with extremes P_0, P_1, with respective coordinates (x_0, y_0) and (x_1, y_1) (Figure 7).

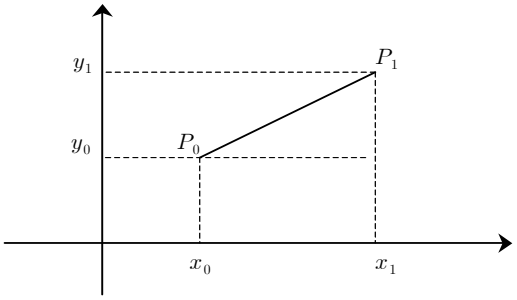


Figure 1.7.

From the Pythagorean theorem we get that the length of the segment P_0P_1, i.e., the distance between the two points P_0 and P_1, is

$$\operatorname{dist}(P_0, P_1) = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2}. \tag{1.15}$$

• *Line passing through two points.* To every straight line in the plane (geometrical entity) it is possible to associate an equation of the type

$$ax + by + c = 0, \tag{1.16}$$

where x and y represent the coordinates of the generic point on the line[9]. Indeed, let (x_0, y_0) and (x_1, y_1) be the coordinates of two points on the line. If $x_0 = x_1$, the line is parallel to the ordinates axis and its equation is simply

$$x = x_0,$$

[9] Which are therefore not to be regarded as unknowns! Here, the actual unknowns are a, b, c.

i.e., of the type (1.16) with $b = 0$. If $y_0 = y_1$, the line is parallel to the abscissae axis and its equation is simply

$$y = y_0,$$

i.e., of the type (1.16) with $a = 0$. If the line is neither "vertical" nor "horizontal", then we have the case depicted in Figure 8.

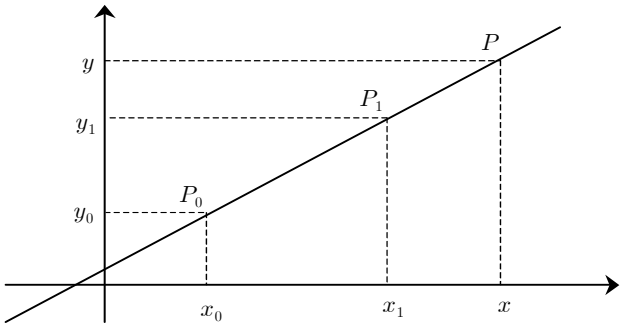


Figure 1.8. Line passing through the points P_0 and P_1

Thales' theorem about a number of parallel lines cut by a transversal yields the equality

$$\frac{y - y_0}{y_1 - y_0} = \frac{x - x_0}{x_1 - x_0},$$

which can be written in the form (1.16). It is more significant to write it as

$$y - y_0 = \frac{y_1 - y_0}{x_1 - x_0}(x - x_0),$$

thus enhancing the coefficient $\dfrac{y_1 - y_0}{x_1 - x_0}$, called the *slope* or the *angular coefficient* of the line, because it determines its slant with respect to the abscissae axis.

Generally speaking, a non-vertical line can be represented by means of an equation of the type $y = mx + q$, where m is the slope. If $m = 0$, we again find the horizontal lines, which thus have null slope. The slope of a vertical line is not defined, but sometimes we shall allow ourselves to say that it is *infinite*.

• *Circumference of radius r.* A circumference with centre at a point C and radius r is the set of those points of the plane whose distance from C is exactly equal to r. This geometrical characterisation easily translates into its corresponding algebraic concept. Once the reference system is fixed, if the centre C and the generic point on the circumference have respectively coordinates (c_1, c_2) and (x, y), thanks to formula (1.15) we simply have to impose that their distance be equal to r:

$$\sqrt{(x - c_1)^2 + (y - c_2)^2} = r.$$

By squaring both sides, we get the algebraic equivalent of our circumference, i.e., the equation

$$(x - c_1)^2 + (y - c_2)^2 = r^2.$$

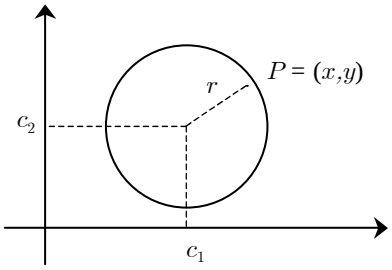


Figure 1.9. Circumference of centre (c_1, c_2) and radius r

• *Parabola.* A parabola can be characterised as the set of those points of a plane which have the same distance from a given line (*directrix*) and a given point (*focus*). As for the circumference, the algebraic object corresponding to a parabola is an equation which links the coordinates (x, y) of its generic point P. To find it, let us fix the reference system in such a way that the focus F has convenient coordinates, for instance $(0, k)$, and the directrix is horizontal with equation $y = -k$.

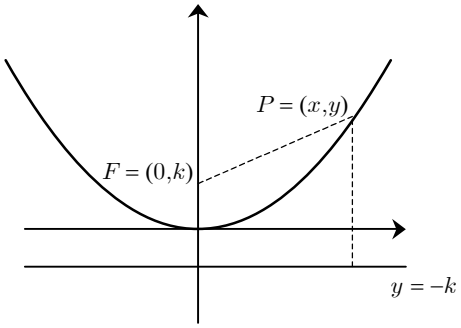


Figure 1.10. Parabola with vertex at the origin and focus at the point $(0, k)$

Let us now translate the condition that a generic point (x, y) is equally distant from the focus and the directrix. A quick glance at Figure 10 reveals that

$$\begin{aligned} \text{dist}\,(P, F) &= \sqrt{x^2 + (y-k)^2} \\ \text{dist}\,(P, d) &= |y + k|\,. \end{aligned}$$

By setting the distances equal and squaring (everything is non negative) we get

$$x^2 + (y-k)^2 = (y+k)^2$$

which leads to the equation

$$y = \frac{1}{4k}x^2.$$

The origin is a special point of such a parabola, called the *vertex*.

1.8 How many elements in a set?

An activity which is common to all human beings from a certain age onwards is certainly that of *counting*. The action of counting (one, two, three...) consists of setting a one-to-one correspondence between a set of objects and the natural numbers $1, 2, 3 \ldots$. In a large part of cases, such a process comes to an end, and the final number determines *the number* of the elements in the set. We also say that the set is *finite*. Of course, if we tried to count all of the elements in the set of the even numbers, we would never stop. Indeed, we call such a set *infinite*.

1.8.1 Finite set. Combinatorics

The problem of determining how many elements there are in a given finite set is elementarily and systematically posed in statistical and probabilistic issues. A typical situation is the evaluation of the *probability* (in the classical meaning) of an event as the ratio

$$\frac{\text{number of cases favourable to the event}}{\text{number of possible cases}}.$$

To evaluate numerator and denominator, we need to "count up", respectively, how many the favourable and the possible cases are. In many cases, the calculation is not trivial at all, and calls for a certain skill. Consider, for instance, the following questions.

(a) Eight pool players are matched against eight other players in games of singles (one against the other). How many are the possible pairings?

(b) How many anagrams are there for the Italian word ANAGRAMMA?

(c) How many ways are there to choose 5 playing cards from a deck of 52?

As a first step, it is advisable to try and understand which kind of sets we are dealing with. In the first case, once the eight players of a team are fixed, all of the matchings can be found by rotating, or better *permutating*, the other eight players in all of the possible ways. In the second case, we are still dealing with permutations, this time among the letters A, G, M, N, R, with the significant difference that some of them are repeated: we then speak of *permutations with repetitions*. Note that the ordering used to list the elements has a relevant role in both cases.

Finally, in the third case, the choice of 5 cards out of 52 is only determined by the chosen cards and not by the sequence in which they are picked: groupings of this kind are called *simple combinations*.

In all of the above situations the point is to settle the number of some particular groupings, built from a given number of different elements. This is what we want to consider when introducing the first elements of *combinatorial calculus*. To proceed further, it is time to delineate some types of groupings by means of precise definitions.

Simple permutations and permutations with repetitions

Let X be a given set of n *different* elements.

Definition 8.1. *Every ordering of n different objects is called a* ***simple permutation***.

For instance, let $X = \{a, b, c\}$. The possible orderings of the three objects a, b, c are six:

$$abc \qquad acb \qquad bac \qquad bca \qquad cab \qquad cba.$$

As it is easy to see, two permutations are distinguished only by the ordering of the elements. Why are they 6? To answer, note that there are three choices for the first place in the listing: a or b or c. However, when the first element has been chosen, just two possibilities for the second are left, and thus we have overall $3 \cdot 2$ choices for the first two elements. Now, for every one of the 6 possible choices for the first two places, there is only one possibility left for the third, and thus the total number for the three places equals $3 \cdot 2 \cdot 1 = 6$. The number $3 \cdot 2 \cdot 1$, the product of the factors from 1 to 3, can be written with the notation 3!, which reads "three factorial". Analogously, we define $6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$ (six factorial) and, generally,

$$n! = n \cdot (n-1) \cdot (n-2) \cdots 3 \cdot 2 \cdot 1,$$

which reads "en factorial". As a convention, $0! = 1$.

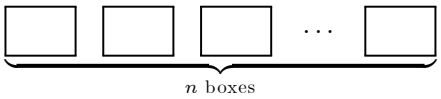
The integer number $n!$ grows very quickly with respect to n. Here are some values:

n	4	5	6	7	8	9	10
$n!$	24	120	720	5040	40320	362880	3 628 800

How many are the permutations of n different objects? Let us denote their number by P_n. As the reader probably already suspects:

Theorem 8.1. $P_n = n!$

Proof. Let us generalise the argument seen above for the permutations of a, b, c. Consider n boxes, as in the figure below.



The number of permutations of the n objects equals the number of ways the n objects can be put one per box.

The first box can be filled in n ways. Once the first is full, there are $n-1$ choices left for the second, and so the first two boxes can be filled in $n(n-1)$ ways. Analogously, for the first three boxes there are $n(n-1)(n-2)$ choices.

When filling the k-th box, $k-1$ objects have been placed, and therefore the box can be filled in $n-(k-1) = n-k+1$ ways. As a consequence, the whole of the first k boxes can be filled in

$$n\,(n-1)\,(n-2)\cdots(n-k+1) \tag{1.17}$$

ways. When $k = n$, we are at the last box and (1.17) equals $n!$. □

We can now answer the first of the questions we asked on page 26: the number of the possible matchings is

$$P_8 = 8! = 40320.$$

To answer the second question, we need to introduce the following

Definition 8.2. *Every ordering of n objects, not all distinct, is called a* ***permutation with repetitions****.*

The word ANAGRAMMA involves a permutation with repetitions of 9 letters (the objects), 4 of which (the As) are equal, 2 more (the Ms) are equal as well, and the other 3 are distinct and different from the previous ones. Without distiguishing the As among themselves and the Ms between themselves, it is clear that the number of distinct permutations is smaller than 9!. To evaluate such a number, let us temporarily label the equal letters and consider the word

$$\mathrm{A}_1\mathrm{NA}_2\mathrm{GRA}_3\mathrm{M}_1\mathrm{M}_2\mathrm{A}_4 \tag{1.18}$$

whose letters are now all distinct. Note now that each one of the $4! = 24$ permutations of the letters A_1, A_2, A_3, A_4, keeping fixed all of the other letters, corresponds to a single permutation of the original word. We can immediately deduce that the permutations which do not distinguish the A letters from one another are $9!/4!$. Analogously, each of the two permutations of the letters M_1, M_2 corresponds to a single permutation of the original word. As a consequence, the number of required anagrams is

$$\frac{9!}{4!2!} = 7560.$$

Generally speaking, let X be a set with n objects such that:

k_1 are equal to one another,

k_2 are equal to one another and distinct from the previous ones,

$\vdots$

k_h are equal to one another and distinct from the previous ones,

with

$$k_1 + k_2 + \cdots + k_h = n.$$

How many distinct permutations of such objects are there? Let us denote their number by $P^*_{k_1,k_2,\ldots,k_h}$. We have:

Theorem 8.2. $P^*_{k_1,k_2,\ldots,k_h} = \dfrac{n!}{k_1!k_2!\cdots k_h!}$.

Simple combinations and binomial coefficients

Among the groupings which do not keep track of the ordering, we have *combinations*. Let us consider again a set of n distinct objects.

Definition 8.3. *Every subset of X with k distinct elements ($0 \le k \le n$) is called a class k* ***simple combination*** *of n objects.*

For instance, out of the set $X = \{a, b, c, d\}$ the following 6 subsets with 2 elements can be extracted:

$$\{a, b\} \qquad \{a, c\} \qquad \{a, d\} \qquad \{b, c\} \qquad \{b, d\} \qquad \{c, d\}\,.$$

Let us denote by $C_{n,k}$ the number of the class k simple combinations of n objects. Thus, $C_{n,k}$ corresponds to the different ways k objects can be chosen out of n distinct given ones. Then:

Theorem 8.3. $C_{n,k} = \dfrac{n!}{k!\,(n-k)!}$.

Keeping in mind that $n! = n \cdot (n-1) \cdots (n-k+1) \cdot (n-k)!$, we can also write

$$C_{n,k} = \frac{n \cdot (n-1) \cdots (n-k+1)}{k!},$$

which is more convenient for calculation purposes.

Proof. Every subset of X with k elements can be identified by applying, say, a white label on k elements and a black label on the remaining $n-k$. This way a permutation is identified, of n elements among which k are equal among themselves, and $n-k$ are equal among themselves and different from the previous ones. Conversely, every labelling which follows the above rules identifies a subset of X with k elements. As a consequence, counting the subsets of X with k elements is equivalent to counting the possible labellings, that is the possible permutations of n elements as above:

$$P^*_{k,n-k} = \frac{n!}{k!\,(n-k)!}. \square$$

Finally, the problem about playing cards can also be solved. We want to pick 5 out of 52 cards, all distinct. This can be accomplished in

$$C_{52,5} = \frac{52!}{5!47!} = \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}{5!} = 2\,598\,960$$

different ways.

The reader is encouraged to find out how many chances there are to guess a tern (three winning numbers) played on the Italian Lotto, or to get a flush while playing *poker*.

• *Binomial coefficients.* The numbers $C_{n,k}$ have gained a particular attention, because they recur in important topics. They indeed have a name, *binomial coefficients*, and are denoted by a special symbol: we set

$$\binom{n}{k} := \frac{n!}{k!\,(n-k)!}.$$

The reason for their name is made clear by the following formula, which is a generalisation of other formulae studied long since, such as the square or the cube of a binomial.

Theorem 8.4 (Newton's binomial). *If a, b are real numbers and n is a natural integer,*

$$\begin{aligned}(a+b)^n &= \sum_{k=0}^{n}\binom{n}{k}a^{n-k}b^k = \\ &= a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \cdots + \binom{n}{n-1}ab^{n-1} + b^n.\end{aligned}$$

We set out the proof of this theorem, as it is interesting in its combinatorial nature.

Proof. Let us write

$$(a+b)^n = \underbrace{(a+b)\,(a+b)\cdots(a+b)}_{n \text{ times}}$$

and note that the rightmost product gets expanded into a sum of degree n monomials of the type $a^{n-k}b^k$, with k ranging from 0 to n. How many monomials with the same k are there? As many as the permutations of n objects (the factors of the monomial), k of which are equal to b and the remaining ones equal to a, that is $\binom{n}{k}$. □

Simple arrangements and arrangements with repetitions

The last type of groupings we want to analyse are *arrangements*. Let X be the usual set of n distinct objects:

Definition 8.4. *Every ordering of k objects, chosen in any possible way among the n ones ($0 \le k \le n$), is called a class k* ***simple arrangement*** *of n objects.*

Two arrangements can thus differ both in the objects they contain and in their ordering. Their number is denoted by $D_{n,k}$. Of course, if $n = k$ we are back to the permutations of n distinct objects, and thus $D_{n,n} = P_n = n!$.

For instance, the class 2 simple arrangements of the 4 objects a, b, c, d are the following 12:

$$ab \quad ba \quad ac \quad ca \quad ad \quad da \quad bc \quad cb \quad bd \quad db \quad cd \quad dc.$$

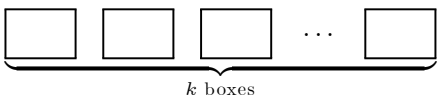
Following the same idea as in the case of simple permutations, we find immediately that:

Theorem 8.5. $D_{n,k} = n\,(n-1)\,(n-2)\cdots(n-k+1) = k!\,C_{n,k}$.

If we allow the elements of an arrangement to be repeated, we are dealing with a class k *arrangement with repetitions* of n objects. In such a case there are no restrictions on k, which is allowed to be even greater than n. The number of such arrangements is denoted by $D^*_{n,k}$. We have:

Theorem 8.6. $D^*_{n,k} = n^k$.

Proof. It is enough to note that, considering the k boxes in the figure



every box can be filled in n different ways, as the elements can be used more than once. □

Typical examples of arrangements with repetitions are the *Totocalcio* (pools) columns, where 3 objects, 1, X, 2, must be placed in 13 distinct cells. Their number is then $D^*_{3,13} = 3^{13} = 1\,594\,323$.

1.8.2 Infinite sets. Countability, power of the continuum

When introducing Set theory, we began with a comment about infinity in Mathematics. We now want to close this section by briefly returning to this topic, and trying to answer, at least at an intuitive level, questions such as "Are there more rational numbers than integers? Are there more real numbers than rational numbers?".

At a first glance, there do not seem to be any problems: since $\mathbb{Z} \subset \mathbb{Q}$, the rational numbers would appear "more" than the integer ones. Analogously, since $\mathbb{Q} \subset \mathbb{R}$, the real numbers seem to be more than the rational ones. Indeed, from the point of view of set inclusion, the argument is correct. But the problem acquires a new and surprising aspect when we think again about the meaning of counting.

Let A and B be two finite sets. To count their elements, they have to be put in one-to-one correspondence with two subsets of the natural numbers, for instance A with $\{1, 2, \cdots, n\}$ and B with $\{1, 2, \cdots, n'\}$. If $n < n'$, one says that the elements in A are less than those in B. We can immediately realise that this is equivalent to the possibility of *setting a one-to-one correspondence between A and a proper subset of B*. Thus, in particular, if $A \subset B$, the elements in A are certainly fewer than those in B. On the other hand, if $n = n'$, we say that the number of elements in the two sets coincide, and this is equivalent to the possibility of *putting A in a one-to-one correspondence with B*.

Definition 8.5. *If two sets can be put in a one-to-one correspondence, they are said to be* ***equipotent****, or to have the same* ***cardinality****, or also to have the same cardinal number of elements.*

When dealing with finite sets, being equipotent, i.e. having the same *cardinal number*, is equivalent to having the same number of elements in the usual meaning, and viceversa. When dealing with *infinite* sets, the problem gets a little tougher. Consider the set $\mathbb{N}$ of the natural numbers and its *proper* subset P of the even integers. Let us have a look at the following scheme.

$$\begin{array}{ccccccc} 0 & 1 & 2 & \cdots & n & \cdots \\ \updownarrow & \updownarrow & \updownarrow & & \updownarrow & \\ 0 & 2 & 4 & \cdots & 2n & \cdots \end{array}$$

The top line contains the elements in $\mathbb{N}$, the bottom one contains those in P. The scheme shows that the two sets are in a one-to-one correspondence, and thus they are *equipotent*, notwithstanding that $P \subset \mathbb{N}$!! With infinite sets, then, more caution is needed, and some further distinctions have to be made. In particular, to decide whether there are more elements in P or in $\mathbb{N}$, it is necessary to specify whether we want to compare the sets from the point of view of the inclusion or from that of cardinality. In the first case we are led to say that $\mathbb{N}$ has more elements than P,

in the second that they have *the same cardinal number of elements*. The sets which are equipotent with $\mathbb{N}$ are said to be *countable*, or to have *countably infinitely many elements*. Such a terminology is due to the fact that the elements in a countable set can be *listed*, or *counted*. It is easy to be convinced (and we invite the reader to do so) that $\mathbb{Z}$ is countable. A little bit more surprising is the fact that $\mathbb{Q}$ is countable as well: $\mathbb{N}$ and $\mathbb{Q}$ are then equipotent. On the contrary, $\mathbb{R}$ is not countable: even from the point of view of the cardinality, $\mathbb{R}$ has *more elements than* $\mathbb{Q}$. To recall that the real numbers are equipotent with the "continuum" of the points in a straight line, we say that $\mathbb{R}$ has the *power of the continuum*. Other sets with the power of the continuum are the intervals we defined before, but also the set of all of the points in a plane or in space.

1.9 Exercises

1.1. Let $a \geq 0$ be a real number such that $a < \dfrac{1}{n}$, for every natural integer n. Then $a =?$

1.2. Suppose that the charged rate for x hours of phone calls in a month is (in Euro)

$$p(x) = x^2 - 80x + 1400.$$

How long can I call if I do not want to spend more than 700 Euro?

1.3. Let $a \neq 0$. The solution of the inequality

$$ax + b < 0$$

is the set of x such that

$$x < -\frac{b}{a}.$$

True or false?

1.4. To solve the inequality

$$\frac{2x}{x^2+1} > 1,$$

multiply both sides by $x^2 + 1$, thus getting

$$2x > x^2 + 1,$$

i.e.,

$$x^2 - 2x + 1 = (x-1)^2 < 0,$$

which is impossible. Is the argument correct?

Consider now the rational inequality

$$\frac{2x}{x^2-1} > 1.$$

and follow the same argument, thus getting

$$2x > x^2 - 1,$$

whence

$$x^2 - 2x - 1 < 0$$

and finally

$$1 - \sqrt{2} < x < 1 + \sqrt{2}.$$

Is this correct?

1.5. Consider the inequalities

$$(a)\sqrt{x^2-5} < 2, \qquad (b)\sqrt{x^2-5} < -2, \qquad (c)\sqrt{x^2-5} > -2$$

By squaring both sides we get, respectively,

$$(a)\ x^2 - 5 < 4, \qquad (b)\ x^2 - 5 < 4, \qquad (c)\ x^2 - 5 > 4$$

whose solutions are

$$(a) -3 < \ x < 3, \qquad (b) -3 < \ x < 3, \qquad (c)\ x < -3 \text{ or } x > 3.$$

Is the argument correct?

1.6. Given two positive numbers a and b, their *arithmetic mean* is $\dfrac{a+b}{2}$, whereas their *geometric mean* is $\sqrt{ab}$. Which of the two is the larger?

1.7. Solve geometrically (that is in terms of distance) the inequalities

$$(a)\ |2x-4| < 1, \qquad (b)\ |x-2| > |x-3|$$

and draw the result on the real line.

1.8. If $\ln x = 2\ln a + 3\ln b - \ln c$, then $x = \ldots$

1.9. To solve the following inequalities

$$(a)\ 3^{x^2} < 81, \qquad (b)\ \log_3\left(x^2-3\right) < 0$$

let us proceed as follows:

(a) since $81 = 3^4$, it is enough that $x^2 < 4$, with solution $-2 < x < 2$;

(b) since every logarithm is positive if the argument is greater than 1, it is enough that $x^2 - 3 < 1$, i.e., $x^2 < 4$, hence the solution is again $-2 < x < 2$.

Any objections?

1.10. Use the summation symbol to write down the following sums:

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots + \frac{1}{900}\,, \qquad 1 - \frac{1}{4} + \frac{1}{9} - \frac{1}{16} + \cdots - \frac{1}{900}.$$

1.11. Show that

$$\sum_{k=1}^{n} (a_k - a_{k-1}) = a_n - a_0.$$

1.12. State whether the following equalities are true or false.

$$(a)\sum_{k=1}^{100} k^3 = \sum_{k=2}^{101}(k+1)^3, \qquad (b)\sum_{k=1}^{100} k^3 = \sum_{k=10}^{110}(k-11)^3.$$

1.13. Is formula (1.8) true also when $c < 0$? Calculate the sum of the first 10 terms in the sequence

$$40,\ 37,\ 34,\ 31, ...$$

1.14. State whether the following equalities are true or false

$$(a)\ \sum_{i=1}^{n}(\alpha x_i + \beta y_i) = \alpha\sum_{i=1}^{n} x_i + \beta\sum_{i=1}^{n} y_i, \quad (b)\ \prod_{i=1}^{n}(ax_i) = a^n\prod_{i=1}^{n} x_i,$$
$$(c)\ \prod_{i=1}^{n}(x_i + y_i) = \prod_{i=1}^{n} x_i + \prod_{i=1}^{n} y_i, \qquad (d)\ \log_a\prod_{i=1}^{n} x_i = \sum_{i=1}^{n}\log_a x_i.$$

1.15. Two firms, both producing hamster seed, are competing against each other. The first produces one metric ton of feed per week and aims at increasing production by 200 kilograms per week. The second starts with half a ton and aims at increasing production by a weekly 20%. Write the formulae describing the weekly production of the two firms for the n-th week and the total production up to then. By means of a computer, draw graphs of the total production in the first 15 weeks.

1.16. Heating a condominium requires, at present, 100 000 litres of oil. An annual increase of either (a) 5000 litres or (b) 5% is forecast (every year). Compute the required fuel amount after 5 years and the total consumption up to then in both cases (a) and (b).

1.17. If $A \subset B$ then $A^c \supset B^c$. True or false?

1.18. The intersection of the two intervals $[2, 7)$ and $(7, 9]$ is:
(a) $(2, 9)$, (b) the empty set, (c) $\{7\}$, (d) meaningless.
Their union is:
(a) $[2, 9]$, (b) $[2, 9] \setminus \{7\}$, (c) meaningless.

1.19. Let r be a straight line with equation $y = 4x + 2001$. What position have the lines with equations

$$y = 4x - \pi \quad \text{and} \quad y = -0.25x$$

with respect to r?

1.20. Prove that the binomial coefficients $\binom{n}{k}$ satisfy the properties

$$\binom{n}{k} = \binom{n}{n-k}$$
$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

1.21. Using the binomial formula, expand $(1+x)^n$, $(1-x)^n$.

1.22. Evaluate the sums

$$\sum_{k=0}^{n} \binom{n}{k}, \qquad \sum_{k=0}^{n} (-1)^k \binom{n}{k}.$$

1.23. If the set X has n elements, how many elements are there in $\mathbb{P}(X)$?

1.24. We want to distinguish 40000 different items in a catalogue by means of an alphanumerical code such as A723C, starting and finishing with a letter and featuring three digits (from 0 to 9) in between. How many codes can be constructed this way?

1.25. A population is divided into three income brackets, respectively containing 10000, 5000 and 1000 families. For the purposes of a market survey, how many ways can 100 families from the first bracket, 50 from the second and 10 from the third be chosen for interviewing? How many ways, then, can the sample of 160 families be composed?

1.26. In a group of 25 people, born in the same (non-leap) year, is it more probable to find at least two people born on the same day (of the same month) or that all of them are born on different days?

2

Functions

≪From this time onward the idea of "function" became fundamental in Analysis.≫

This sentence from the book by C. B. Boyer, *A History of Mathematics*[1], refers to the publication of the two volumes of Euler's *Introductio in analysin infinitorum*, in 1748. Our introductory sentence points out that the idea of *function* is the result of long meditations and debates among the most famous scientists of the seventeenth and eighteenth centuries, such as Euler, Newton[2], Leibniz[3], d'Alembert (1717-1783), Daniel Bernoulli (1700-1782).

The term *function* is used in many fields of knowledge and in a great variety of everyday situations, with different specific meanings. In sentences like: "so-and-so has taken on the *function* of sales management" or "the *function* of the heart is to pump blood", we note that the term *function* suggests a role or a purpose. In another context, like: "to attend a (social or sacred) *function*" it takes the meaning of a ceremony or ritual.

In the previous meanings everyone should be able to notice the presence of a dynamic common nuance, which we could regard as a *cause-effect* implication, with an emphasis mainly or exclusively on the *effect*. In biology, for instance, when we talk about the physiological function of an organ, we underline its role for the fulfilment of a specific purpose, rather than the *procedure* for its achievement.

---

[1] John Wiley & Sons, New York, 1968.

[2] Sir Isaac Newton (1643-1727), English physicist and mathematician, who invented infinitesimal calculus. Newton and Leibniz achieved the same result, each one independently and separately from the other.

[3] Gottfried Wilhelm von Leibniz (1646-1716), German mathematician and philosopher, who was also interested in financial and actuarial calculus, although with less successful results.

In Mathematics, as well as in Physics, Economics and Chemistry, the idea of *function* has developed trying to capture the *whole* dynamics, not only the final effect.

After defining constants (numerical parameters, in practise) and variables (such as time, space...), according to Euler: ≪a function depending on a variable is an analytic expression, composed in whatever way of that variable quantity, and of numbers or constant quantities≫.

Clearly and for the first time, he based the idea of function on the *dependence of one quantity on another, considered as variable*, and the stress is on the analytic expression, that is *on how* the two quantities are related.

Today, the notion of function finds its natural place in set theory and it is no longer connected only with numerical variables. In this chapter we introduce the general concept of function, and then we concentrate our attention on the most important one-variable numerical functions, according to the following outline.

• The general concept of function. In particular, we shall consider:
— functions defined on the set of natural numbers (sequences);
— linear and quadratic functions;
— composite and inverse functions;
— general properties of graphs, such as monotonicity and convexity.

• The main elementary functions: power functions, exponentials, logarithms and trigonometric functions.

2.1 The concept of function

"Temperature varies with altitude", "the cost of goods depends on the amount purchased", "consumption depends on income".

In all these sentences we implicitly use the concept of function. What do we mean by saying, for instance, that "the price of rice is a function of how much rice is supplied"?

We mean that there is a "law" (the "analytic expression" in Euler's statement) which allows us to deduce the market price of rice, given its supplied quantity: during years of plenty, there will be a high supply and sellers will keep the price low, in order to sell all their produce. The opposite will obviously happen in years of under-production. The following diagram illustrates the underlying dynamics of this dependence

$$\text{supply} \longrightarrow \boxed{\textbf{function}} \longrightarrow \text{price}$$

where the function looks like a "*black box*", which associates with each *input* x exactly one *output* y :

$$\underset{\text{input}}{x} \longrightarrow \boxed{\textbf{black box}} \longrightarrow \underset{\text{output}}{y}$$

Definition 1.1. *Given two sets A and B, a* ***function*** *from A to B is a law that associates with each element of A one (single) element of B. The set A is called the domain of the function, while the set B is called the codomain.*

It is worth emphasizing an essential feature of the definition: given an *input,* the *output* must be *unique.*

As one can see, there are no restrictions on the choice of sets A and B. Let us consider some examples.

Examples

1.1. If A is the set of polygons in a plane and B is the set of real numbers, the correspondence

$$\text{polygon} \mapsto \text{area}$$

is a function from A into B.

1.2. The correspondence

$$\text{son} \mapsto \text{mother}$$

turns out to be a function from the set U of human beings to the set of women D. We note that the mapping *mother* $\mapsto$ *son* is not a function: every parent may actually have more than one son.

1.3. The correspondence

$$\text{apple} \mapsto \text{tree (to which the apple belongs)}$$

is a function from the set D of apples to the set A of trees.

We usually refer to functions by the letters f, g, h, F, G... and we write

$$f : A \to B$$

(to be read "f from A into B"). If the function associates with the element $x \in A$ (*input*) the element $y \in B$ (*output*) we write $f : x \mapsto y$ (to be read "f maps x into y").

The element y, denoted also by the symbol $f(x)$ (to be read "f of x"), is called the *image* of x under f, while x is an *inverse image* of y.

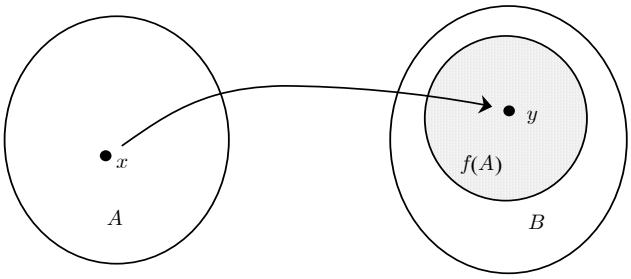


Figure 2.1. Function from A to B

The set of all possible *outputs* coming from all possible *inputs* in A is called the *range* or *image* of A under f; it is denoted by the symbol[4] $f(A)$. The *input* variable

[4]We note that the range can either coincide with the codomain or be a proper subset of the codomain: $f(A) \subseteq B$. .

in set A, denoted by x (or by any letter different from f, A or B...), is called the *independent variable*, the *output* variable in set B is said to be the *dependent variable*.

The above examples are not really interesting in Mathematics. We are interested in "numerical" functions, where *input* and *output variables* are numbers or groups of numbers. In this chapter real numbers come into play as variables, and for this reason we will talk about *real functions of a real variable.* Examples of this kind of functions are: the production cost of goods as a function of the produced amount, the production of (barrels of) oil in an Arabian country as a function of time, a car's speed as a function of its position. Therefore, in all these cases we are dealing with laws that associate with a real number x *one* real number y *only*, so that they have a subset A of $\mathbb{R}$ as domain and $\mathbb{R}$ as codomain.

Studying the *graph* of a function in an appropriate Cartesian plane turns out to be very useful, in order to visualize its behaviour[5].

Definition 1.2. *The* ***graph*** *of a function $f : A \to B$ is the set of pairs (x, y), with $x \in A$, $y \in B$, such that $y = f(x)$.*

In most cases, the graph of a real function of a real variable is a plane curve. It is very important to keep in mind the graphs of all the frequently used functions, referred to in this chapter.

Two functions are equal if they have the same graph; this implies that they have the same domain A and the same range $f(A)$. We have pointed out that the peculiar feature characterising a function is that to a single *input* must correspond one *output* only. This feature can be translated into the following geometric condition:

each vertical line meets the graph of f at most at one point .

In other words, either of the following conditions holds: the vertical line does not intersect the graph or it intersects the graph exactly at one point. Therefore not all curves in the plane are graphs of functions. For instance, the fact that a vertical line can cross a circumference in two points tells us that this curve cannot be the graph of a function: two different values of y (coordinates on the y-axis of the intersection points) correspond to the value x defining the vertical line (figure 2), invalidating the claim that to every x in the domain can be associated *only one* y in the codomain.

When the law defining the correspondence between the elements of A and B is given by an equation in two variables x and y, for instance in the form $y = f(x)$, we take for granted (unless otherwise stated) that the domain of f is the largest subset of $\mathbb{R}$ where all operations indicated in the expression $f(x)$ can be performed. This subset is called the *natural domain*[6] of f. When any information about the codomain is omitted, it is understood to be $\mathbb{R}$.

Examples

1.4. The natural domain of $f(x) = 2x^3 - 5x^2 + 7$ is the whole set $\mathbb{R}$.

---

[5] It seems that the first person to use graphs in a systematic way was Nicole d'Oresme (1320-1382), bishop of Lisieux, in Normandy.

[6] Some authors call it the *existence set* or *definition set.*

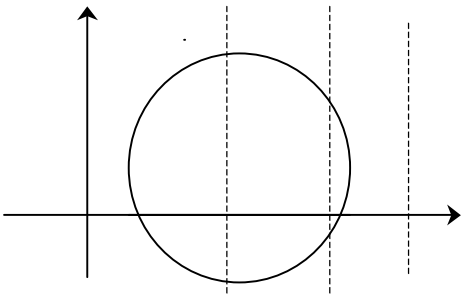


Figure 2.2. A circumference is not the graph of a function

1.5. The natural domain of $f(x) = \dfrac{1}{x-3}$ is $\mathbb{R}$ with a "hole" at point 3, that is $\mathbb{R} \setminus \{3\}$.

1.6. The natural domain of $f(x) = \sqrt{2-x}$ is the interval $(-\infty, 2]$, since the square root is only defined for $2 - x \geq 0$.

We note that when we use functions in order to build models for applications, we might be interested not in the whole natural domain, but in one of its parts only. Consider the case of the function $f(x) = 1000/x$, telling us what the unitary price $f(x)$ of goods would be if we want to sell the quantity x. It is clear that the natural domain of f is $\mathbb{R}$ with a "hole at the origin", that is $\mathbb{R} \setminus \{0\}$ and that we could compute $f(x)$ also for $x < 0$, but it is obvious that the result would not have any interesting economic meaning. We therefore have to be careful: the natural domain of a function may be too wide for some of its uses.

It is meaningful to use real numbers to measure quantities mainly when those quantities "vary continuously". It is, then, natural to think of real variables when describing the amount of fluid in a tank, the space covered by an object moving along a line, or through time...

For other quantities a *discrete* variable description seems more natural or more convenient, when integer values are assumed like, for instance, the quantity of some goods[7]. A firm producing biscuits will count the number of boxes produced, a firm producing wine the number of bottles etc. In these cases quantities are "measured" by natural integers. Models are described by *sequences*, which are functions defined on $\mathbb{N}$.

In some cases time can also be considered as a *discrete variable*. Studying mathematical models for economic systems, sequences appear in the case of periodic observations. Choosing the time interval between two subsequent observations as the unit of measure of time, the time variable assumes only natural values.

[7] It is meaningless to talk about the cost of $\pi + \sqrt{3}$ cars...

2.2 Sequences

The term *sequence* brings to mind a set whose elements can be put in a list, one after the other. Listing the elements in the set means that a place is given to each object, so that they all have a *natural number as a label.* The element in the position n of the list is called a_n and the law associating a_n with each natural number n is denoted by the symbol

$$n \longmapsto a_n.$$

We say that a_n is the *general term* of the sequence and that n is the *index.* The set made up of the terms

$$a_0,\ a_1,\ a_2,\ a_3, \ldots,\ a_n, \ldots$$

is denoted by the symbol $\{a_n\}$ and is called a *sequence.* We are especially interested in the case when $\{a_n\}$ is a subset of $\mathbb{R}$.

Definition 2.1. *A real-valued* ***sequence*** *is any law associating with each natural integer a real number, that is any function with domain* $\mathbb{N}$ *and codomain* $\mathbb{R}$.

Frequently, a real-valued sequence may be described by means of explicit (or *closed*) formulae like

$$a_n = \frac{1}{n+1}, \qquad \text{or} \qquad b_n = \left(1 + \frac{1}{n}\right)^n.$$

Note that $\{b_n\}$ is defined only from $n = 1$ onwards. This is not a problem: we may count starting from 0, 1, but also from 2004 or from one million... The sequence domain might be a subset of $\mathbb{N}$, provided that it includes all natural numbers from a fixed n_0 onwards.

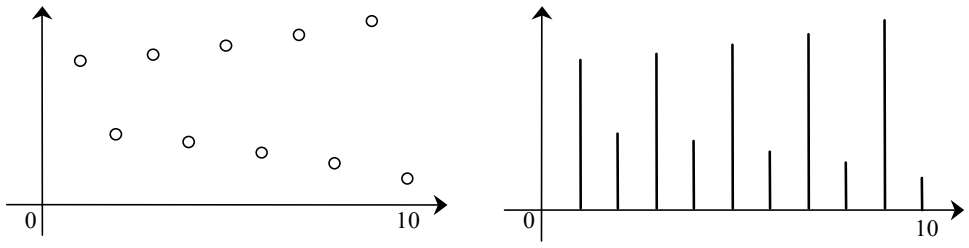


Figure 2.3. Graphic representations of a sequence

If we want to give a geometric representation of a sequence $\{a_n\}$ on a Cartesian plane, we put the values of the index $n = 0, 1, 2, 3, \ldots$ on the x-axis, while the corresponding values $a_0, a_1, a_2, a_3, \ldots$ are put on the y-axis. We obtain a sequence of points with coordinates (n, a_n) forming the *graph* of the sequence. Sometimes the position of the terms a_n may be better visualised by means of a "stick" from the point $(n, 0)$ to (n, a_n), as in figure 3 on the right.

2.2.1 Recursive sequences

Sequences may be defined not only by closed formulae, but also *by recursion*, i.e. through a *recursive* procedure. This procedure, in the simplest cases, consists of two steps:

— a starting value a_0 is assigned (the *initial value*);

— a (so-called *recursive*) law is given, so that a_{n+1} may be computed, once a_n is known. The computation procedure for such a sequence is illustrated in figure 4 and turns out to be particularly suitable for automatic calculations.

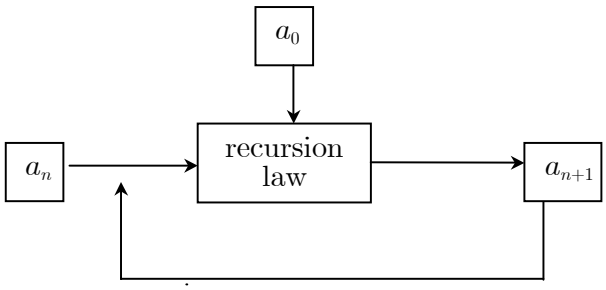


Figure 2.4. Computation loop for a recursive sequence

For example, the *arithmetic sequence* with a as the first term and arithmetic ratio r is defined by recursion as follows

$$\begin{cases} a_0 = a \\ a_{n+1} = a_n + r \end{cases}$$

or explicitly by the formula

$$a_n = a + nr.$$

• (⇛ **Chapter 11**) *Simple interest.* The amount C is invested at simple interest for n years. Let the interest rate of the investment be $r > 0$. The interests earned every year are

$$C \cdot r \cdot 1 = Cr,$$

so that, after n years, the amount of money available (called *maturity value* or *future value*) turns out to be

$$a_n = C + Crn = C\,(1 + rn)\,.$$

We have thus defined a law associating with each year the corresponding final value of the investment

$$n \longmapsto a_n = C\,(1 + rn)\,.$$

The sequence may be defined by recursion. We can state that, starting from an initial capital C, every following year we get a final value that can be obtained by adding the interest paid in the last year (Cr) to the amount already acquired in the previous year. The sequence $\{a_n\}$ is thus defined in a recursive way by

$$\begin{cases} a_0 = C \\ a_{n+1} = a_n + Cr. \end{cases}$$

2.2.2 Geometric sequences

In economic applications we often find sequences $\{a_n\}$ where the ratio between two consecutive terms is constant. Sequences of this kind are called *geometric* (or *exponential*) sequences and the constant value of the ratio between consecutive terms is called the *geometric ratio*. We can represent the general term of a geometric sequence as

$$a_n = aq^n,$$

where the first term is $a_0 = a \neq 0$ (in order to avoid trivialities) and q is the ratio. The same sequence may be defined in a recursive way putting

$$\begin{cases} a_0 = a \\ a_{n+1} = a_n q. \end{cases}$$

The important feature of a geometric sequence is an obvious consequence of the structure of its general term: in such a sequence the *percentage variation* between consecutive terms is always the same. It is actually

$$\frac{a_{n+1} - a_n}{a_n} = \frac{aq^{n+1} - aq^n}{aq^n} = q - 1. \tag{2.1}$$

The ratio q may also be called the *variation rate* and $p = q - 1$ the *percentage variation rate* between two consecutive terms. The general term of a geometric sequence may be written by means of the percentage variation rate:

$$a_n = a\,(1 + p)^n\,. \tag{2.2}$$

Some possible interpretations of the terms of (2.2) are the following:

— a_n = the total population growing (if $p > 0$) or diminishing (if $-1 < p < 0$) for n years, p being the percentage variation rate (a = initial population).

— a_n = a firm's total sales, where the invoice at year zero is a, and sales increase regularly in time, according to the percentage p.

— a_n = the gross domestic product for year n in an economic system with growth rate p, constant in time, having at the year zero a g.d.p. equal to a.

• (⇛ **Chapter 11**) *Compound interest.* Think of an amount C, invested for n years at the interest rate r. Suppose that the capital, after the first year, is reinvested and that the interests (Cr) are added to the capital so that this amount produces interests as well. At the end of the first year the final amount (*future value*) will be $C_1 = C(1 + r)$.

Suppose that this happens for n years and let C_n be the future value at the end of the n-th year. At the end of the second year we shall have

$$C_2 = C_1 + C_1 r = C_1\,(1 + r) = C(1 + r)^2$$

and, in general,

$$C_n = C_{n-1} + C_{n-1}r = C_{n-1}(1+r) = C(1+r)^n.$$

We get a geometric sequence with ratio $1+r$.

2.3 Linear functions

The cost and the amount of purchased goods are usually connected by the most simple relation. Let y be the cost (in Euro), x the purchased quantity (in hectograms, for instance), and m the price for each hectogram, then

$$y = mx. \tag{2.3}$$

This law, called the *law of direct proportionality* , is materially achieved by weighing scales, as we know very well. When the shopkeeper inserts the unitary price for one hectogram, he chooses m; when he puts the goods on the scale pan and thus fixes x, the display immediately shows $y = mx$.

From the law (2.3) it is clear that two variables x and y are directly proportional if *their ratio is constant* ($y/x = m$). There is however another important characterizing property, i.e. the sum and the multiple of proportional quantities keep the same proportionality: in fact, if $y_1 = mx_1$ and $y_2 = mx_2$, we have $y_1 + y_2 = m(x_1 + x_2)$ and if $y = m\ x$, we have $ay = m\,(ax)$. This property characterizes *linear functions.*

Definition 3.1. *A function $f : \mathbb{R} \to \mathbb{R}$ is said to be* ***linear*** *if for all x_1, x_2, x, $a \in \mathbb{R}$, we have*

$$\begin{aligned} f(x_1 + x_2) &= f(x_1) + f(x_2) \quad (f \text{ is additive}) \\ f(ax) &= af(x) \qquad\quad (f \text{ is homogeneous}). \end{aligned}$$

Now, linear functions from $\mathbb{R}$ to $\mathbb{R}$ are exactly functions of type (2.3). In fact:

Theorem 3.1. *A function $f : \mathbb{R} \to \mathbb{R}$ is linear if and only if it is of the type*

$$\boxed{f(x) = mx} \qquad m \in \mathbb{R}. \tag{2.4}$$

Proof. We have already seen that $f(x) = mx$ is linear. Conversely, we try to persuade ourselves that if f is even only homogeneous, it is of the type (2.4). It is sufficient to note that

$$f(x) = f(x \cdot 1) = xf(1),$$

and therefore, putting $f(1) = m$, we obtain (2.4). □

The graph of (2.3), if we think of x and y as real variables, is represented on the plane by a straight line through the origin. The number m is called the line's *slope.* If we use the same unit of measure on both axes, m is connected with the angle α formed by the straight line with the positive half of the x-axis by the relation

$$m = \tan\alpha$$

where $\tan\alpha$ denotes the *trigonometric tangent* of α (this topic will be referred to later).

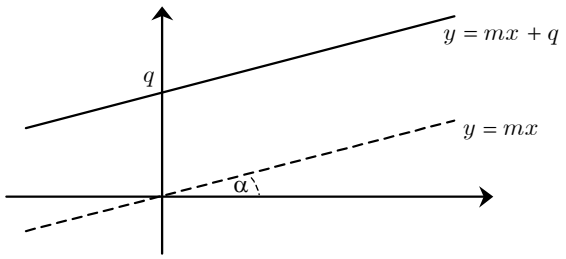


Figure 2.5. Linear and affine linear functions

Affine linear functions

Inappropriately, sometimes the term linear is extended to functions defined by

$$f(x) = mx + q,$$

whose graph is a non-vertical straight line, obtained by shifting the graph of the linear function $f : x \to mx$ up by q, if q is positive, or down by $-q$ units if q is negative. The number $q = f(0)$ is called the *vertical intercept*. The correct name for these functions is *affine* or *affine linear functions*.

• *Market equilibrium.* The demand and supply of goods depend on the price of those goods. In a simple model describing the market price formation, we can assume that demanded and supplied quantities are affine linear functions of the price.

Let q_{d} be the demanded quantity, q_{s} the supplied quantity, p the price, a, b, c, d positive parameters [8]. We have

$$q_{\mathrm{d}}(p) = a - bp, \qquad q_{\mathrm{s}}(p) = -c + dp.$$

The demand *decreases* with increasing price, while supply *is increasing*. The market is in equilibrium if $q_{\mathrm{d}} = q_{\mathrm{s}}$, this happens if

$$a - bp = -c + dp$$

whence we get the *equilibrium* price p^*

$$p^* = \frac{a + c}{b + d}.$$

• *Production cost.* The *costs* met by a firm when producing goods may be divided into: *fixed*, independent of the quantity of the produced goods (connected, for instance, to existing machinery) and *variable*, whose amount depends on the production volume (for example: the purchase of raw material).

If we denote by C the total production cost, by q the production quantity, by f the fixed costs and by $V(q)$ the variable costs, we have

$$C(q) = f + V(q).$$

[8] Parameter a represents the *potential market* . Parameter c may also be zero.

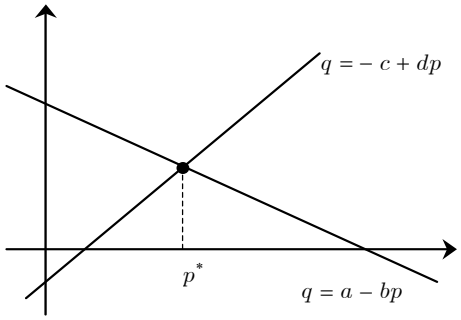


Figure 2.6. Equilibrium price

Let us suppose that the variable costs are directly proportional to the production quantity; if we denote the (constant) unitary variable cost by v, we get the very simple model

$$C(q) = f + vq.$$

The function C is affine linear.

In the case of *perfect competition* the firm's "weight" in the market is not sufficient to have an influence on the goods' sale price p (constant) with the volume of its supply. The total revenue

$$R(q) = pq$$

is a linear function of the amount produced. Suppose that the price covers at least the variable costs, i.e. $p > v$. The profit

$$\pi(q) = R(q) - C(q) = pq - (f + vq) = (p - v)\, q - f$$

is an affine linear function of q. The slope of this straight line is the difference between price and unitary variable cost and is called the *unitary marginal contribution*. The vertical intercept is the fixed cost (with negative sign).

The producing power of our firm will allow it to choose the production volume q in a given acceptable interval, of the type $[q_0, q_1]$ (not less than q_0 and not more than q_1). An interesting question is whether production volumes may be achieved in order to lead the firm to make a profit: in other words, such that revenues cover not only variable, but also fixed costs. Our model immediately gives the answer: the profit is positive if (and only if)

$$\underbrace{(p - v)\, q - f}_{\text{profit}} > 0 \quad \Rightarrow \quad q > \underbrace{\frac{f}{p - v}}_{\text{break-even point}}.$$

This formula shows that the chance of achieving profits is connected to the chance of expanding production beyond a given threshold, called the *break-even point.* As we have already seen, $p - v$ is the unitary contribution margin, that is how much is left by each unit which is produced and sold to cover the fixed costs, therefore the ratio $f/(p - v)$ marks the production level needed to cover f.

• *The indifference point.* A firm's production may be achieved by choosing one of two possible processes, differing in both fixed and unitary variable costs. For the technology of the first line the fixed costs are equal to 450000 Euro and unitary variable costs are equal to 350 Euro, while for the second the fixed costs are 650000 Euro and unitary variable costs are 250 Euro. Since the fixed costs of the second technology are greater, it is obvious that this line is profitable only beyond a minimum threshold of production. Let us compute this threshold. We write the (total) cost functions in both cases, denoting by q the quantity of goods produced

$$\begin{aligned} C_1(q) &= 350q + 450000 \\ C_2(q) &= 250q + 650000. \end{aligned}$$

Putting $C_1 = C_2$, we find

$$\begin{aligned} 350q + 450000 &= 250q + 650000 \\ q^* &= 2000. \end{aligned}$$

Therefore the second line is less expensive for more than 2000 units. We say that q^* is the *indifference point* or the *point of profitability reversal.*

2.4 Quadratic and inverse proportionality

2.4.1 Quadratic functions

A supermarket chain is setting up some sales points. The shape of the commercial surfaces is rectangular and the ratio between the two straight sides has been chosen to follow the rule of the *golden section*, known among ancient Greeks for its pleasantness. Such a ratio is

$$\frac{\text{short side}}{\text{long side}} = \frac{\sqrt{5}-1}{2} \simeq 0.61803$$

and will be denoted by a for short. The measure x of the long side, the *caliber* of the sales point, is sufficient to find the building dimension, by means of the area of the commercial surface ax^2. We have a first example of *quadratic proportionality*: the commercial area is proportional to the square of the caliber.

Many calculations concerning the installation and the management of sales points are based on the hypothesis that a lot of quantities concerning a sales point are proportional to its area. For example, the expected volume of revenue $f(x)$ from a sales point of caliber x is

$$f(x) = b \cdot ax^2 = Ax^2 \quad \text{with} \quad A = ba.$$

With the new constant A, we are dealing again with a quadratic proportionality law.

The graphs of functions

$$\boxed{f(x) = ax^2} \qquad a \neq 0,$$

are represented by *parabolae* having their vertex at the origin and the y-axis as the axis of symmetry. Generalising a bit, the graph of functions $f : \mathbb{R} \to \mathbb{R}$ given by the law

$$\boxed{f(x) = ax^2 + bx + c} \qquad a \neq 0,\ b, c \in \mathbb{R}, \tag{2.5}$$

is a *parabola* with the vertical straight line of equation $x = -\dfrac{b}{2a}$ as the axis of symmetry and with vertex at the point[9] $V = \left(-\dfrac{b}{2a}, -\dfrac{b^2 - 4ac}{4a}\right)$. At the vertex, the function attains its maximum value $-\dfrac{b^2 - 4ac}{4a}$ if $a < 0$, or its minimum value if $a > 0$.

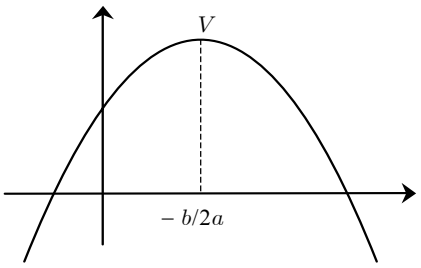


Figure 2.7. Parabola of equation $y = ax^2 + bx + c$

Quadratic functions like

$$\pi(q) = aq^2 + bq + c, \qquad a, c < 0,\ b > 0$$

are suitable for describing a firm's profit as a function of the quantity of produced goods. If the produced quantity is "small", we suppose that the profit may even be negative (there are fixed costs). In order for the firm to make a profit, a certain level of production should be reached. We suppose, then, that the profit increases with the growth of production. Clearly, anyway, it could not increase indefinitely: once the optimum level is reached it will start to decrease, because placing more and more goods in the market causes a progressive and significant revision of the price list.

• *The monopolist*. In a monopoly regime, the sale price of produced goods may be decided by the producing firm. The quantity of goods that the firm will be able to sell depends however on the offered price: if one wants to sell a high quantity it is obviously necessary to offer low sales prices.

Let us denote by $q(p)$ the units of goods sold as a function of the offered price p and suppose that q is affine linear. The function q can be determined by the minimum

[9] This can be seen writing the equation in the form

$$y = ax^2 + bx + c = a\left(x + \frac{b}{2a}\right)^2 + \frac{4ac - b^2}{4a}.$$

applicable price P which reduces sales to zero and by the maximum volume of sales Q, feasible only when the goods are given as gifts (the so-called *potential market*). Given an estimate of P and Q, the (affine linear) function is completely known:

$$q(p) = Q\left(1 - \frac{p}{P}\right).$$

In the simplest case, production cost varies in an affine linear way with the quantity produced and, as a consequence, considering the linearity of q, it turns out to be an affine linear function of price. From

$$C(q) = f + vq$$

we get

$$C\left[q(p)\right] = f + vQ\left(1 - \frac{p}{P}\right) = f + vQ - \frac{vQ}{P}p.$$

Thus we can construct the profit $\pi\left(p\right)$ as a function of all possible choices of the price p

$$\pi(p) = pq(p) - C(q(p)) = -\frac{Q}{P}p^2 + \left(Q + \frac{vQ}{P}\right)p - f - vQ.$$

If we deduce the price as a function of quantity,

$$p(q) = P - \frac{P}{Q}q,$$

we can also construct the function where the profit depends on the quantity which is produced and sold:

$$\pi(q) = p(q)q - C(q) = -\frac{P}{Q}q^2 + (P - v)q - f.$$

2.4.2 Inverse proportionality

Besides the direct proportionality $y = mx$ and quadratic proportionality $y = ax^2$ laws, we also recall the inverse proportionality law. For example, base and height in rectangles with the same area are inversely proportional, as are the quantity of money and its circulation speed according to the so-called "Fisher equation"[10].

Two non-zero variables x and y are inversely proportional if their product is constant. The inverse proportionality law is then

$$xy = k \qquad k \neq 0,$$

which may also be written as $y = \dfrac{k}{x}$. The graph of functions $f : \mathbb{R} \setminus \{0\} \to \mathbb{R}$ defined by

$$f(x) = \frac{k}{x}, \qquad k \neq 0$$

is a *hyperbola* with the Cartesian axes as *asymptotes*.

[10] It generated an enormous literature starting from I. Fisher (1911).

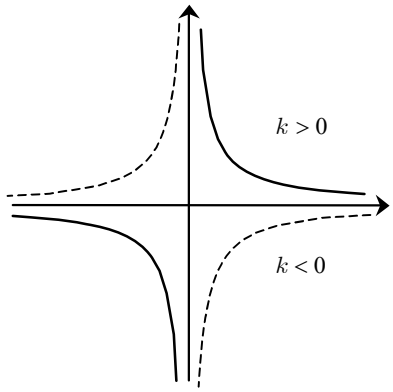


Figure 2.8. Hyperbolas of equation $y = k/x$

For large values of x, the graph of f increasingly approaches the x-axis (horizontal asymptote). Speeding things up a bit, we say that *f tends to zero as x tends to $+\infty$.*

Similarly, if we choose values of x closer and closer to zero, f assumes larger and larger (absolute) values. In this case we might say that *f tends to $\pm\infty$ as x tends to zero.*

• *The demand function.* In economic models, inverse proportionality occurs and has an important meaning. Let p be the sale price of some merchandise and let $q(p)$ be the demanded quantity, usually called the *demand function.* If the seller raises the price, the demanded quantity reduces and the effect on the revenue is generally not predictable: we are selling less, but the price is higher. In the case of an "inversely proportional" demand function, that is $q(p) = k/p$ $(k > 0)$, the effect on the revenue is very simple: it remains unchanged, in fact

$$\text{revenue } = pq(p) = p\frac{k}{p} = k.$$

Such a demand function is called a "constant area of expense" function, because the rectangles giving the revenue have the same area for each price level. In Economics we might say that the elasticity of this function is equal to one.

2.5 Composite function. Inverse function

2.5.1 Composite function

On page 49, studying monopolistic behaviour, we saw that if we substitute the expression for the quantity to be produced as a function of price, $q(p)$, in the total cost function $C(q) = f + vq$, we can directly express the production cost as a function of the offered price $C[q(p)]$. This is a first example of *composite function.* The profit as a function of price was also obtained in the same way, applying one function after the other.

Consider two functions $f : A \to \mathbb{R}$ and $g : B \to \mathbb{R}$, such that *each* value taken by f is in the domain of g (in short $f(A) \subseteq B$). With each $x \in A$, f associates one and only one element $f(x)$ and, since this is an element of the domain of g, the function g associates with it a unique element y, given by $g[f(x)]$. Thus we have the definition of a function $h : A \to \mathbb{R}$, called the *composite* of f and g and denoted by the symbol $g \circ f$. It can be read "g composite with f" or "g circle f".

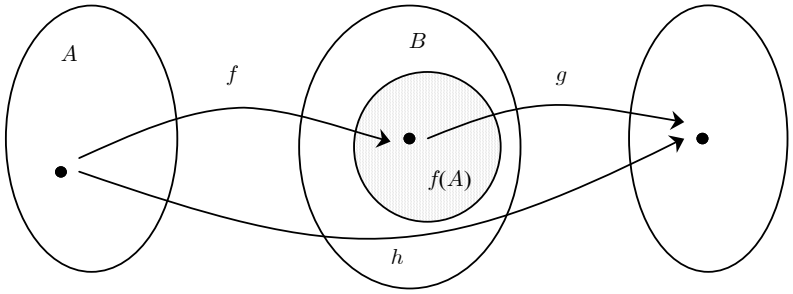


Figure 2.9. The composite function

Examples

5.1. Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be given by

$$f(x) = x + 3, \qquad g(x) = x^3.$$

Since both functions have $\mathbb{R}$ as domain and range, we can construct the two composite functions $g \circ f$ and $f \circ g$. We get

$$(g \circ f)(x) = g[f(x)] = (x+3)^3, \qquad (f \circ g)(x) = f[g(x)] = x^3 + 3.$$

As the example clearly shows, composition does not commute. It may also happen that $g \circ f$ is well defined while $f \circ g$ is not.

5.2. Let $f : (0, +\infty) \to (0, +\infty)$ and $g : \mathbb{R} \to (-\infty, 0]$ be given by

$$f(x) = \frac{1}{\sqrt{x}}, \qquad g(x) = -x^2.$$

Since g is defined on the whole real axis (and thus it includes all values taken by f) the composite function $g \circ f$ is well defined, its domain is the interval $(0, +\infty)$ and we have $(g \circ f)(x) = -1/x$. On the other hand, values taken by g are negative and cannot be used to compute f. In formal terms, this means that the range of g has no common point with the domain of f, thus the function $f \circ g$ cannot be defined.

Composition may be extended to the case of more than two functions: $f \circ g \circ h...$ It satisfies the associative law

$$(f \circ g) \circ h = f \circ (g \circ h).$$

In some cases a function may be composed with itself. In order to denote the function $f \circ f$ we shall use the symbol f^2 (called the *second iteration* of f).

Warning! Do not to confuse $f^2(x)$ with $[f(x)]^2$!

For example, if $f(x) = 1/x$ we have $f^2(x) = x$, while $[f(x)]^2 = 1/x^2$.

2.5.2 Inverse function

In the example of the monopolist (page 49), we constructed the function associating the quantities of goods sold with the offered prices:

$$q(p) = Q - \frac{Q}{P}p. \tag{2.6}$$

Then, from equation (2.6), we deduced the price as a function of the quantity sold

$$p(q) = P - \frac{P}{Q}q.$$

The two functions q and p are each one the inverse of the other. If $q = q(p)$ maps the interval $[0, P]$ into the interval $[0, Q]$, the inverse function $p = p(q)$ maps $[0, Q]$ into $[0, P]$.

The inverse relation of a function f, that is the correspondence associating with the elements in the range of f the respective inverse images, is not always a function. Such a correspondence is a function if and only if each element of the range has exactly one inverse image. In this case f is said to be a *one-to-one correspondence* between the domain A and $f(A)$.

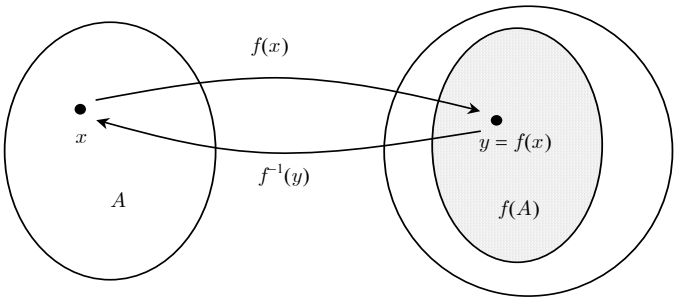


Figure 2.10. A function and its inverse

Definition 5.1. *A function $f : A \subseteq \mathbb{R} \to \mathbb{R}$ is said to be* ***invertible*** *if and only if it is a one-to-one correspondence between A and $f(A)$. The function associating with each $y \in f(A)$ the unique element x such that $f(x) = y$ is called the* ***inverse function*** *of f and it is denoted by the symbol f^{-1}.*

Composing in the two possible ways the couple f, f^{-1} we find the two following relations which characterize the connection between a function and its inverse:

$$f^{-1}[f(x)] = x \qquad f\left[f^{-1}(y)\right] = y.$$

They are valid, respectively, for each $x \in A$ and for each $y \in f(A)$ (in other words, composing a function with its inverse we get the identity function).

The graph of an invertible function can be easily recognized because it is intersected at most once by each horizontal line. The graphs of functions $y = f(x)$ and $y = f^{-1}(x)$ are symmetric with each other with respect to the bisector of the first and third quadrants.

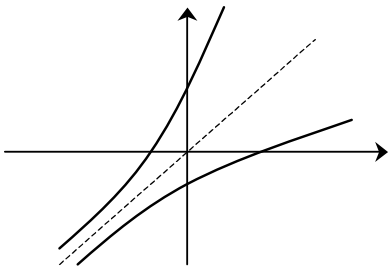


Figure 2.11. Graphs of f and f^{-1}

Knowing the analytic expression of an invertible function f, if we want to find the analytic expression of f^{-1} we have to solve the equation $f(x) = y$ with respect to x.

For example, let $f(x) = 3x - 4$. The equation $3x - 4 = y$ gives $x = (y + 4)/3$. Then, if we want to draw the graphs of f and f^{-1} in the same cartesian coordinate system, we can simply exchange the variables x and y in the equation. Thus we get $f^{-1}(x) = (x + 4)/3$.

As the reader should have guessed, the equation $f(x) = y$ may rarely be solved by means of explicit formulae; then, even if f is known, the analytic expression of f^{-1} can be determined in very few cases.

2.6 Monotonic, bounded, convex functions

In the previous sections we used terms like increasing function, maximum and minimum value, relying on their intuitive meaning. It is time to define these and other concepts concerning the graphical behaviour of functions.

2.6.1 Bounded functions

If the whole graph of a function f, defined on a domain A, lies under some horizontal line of equation $y = K$, the function is said to be *bounded above.* It means that

$$f(x) \leq K,$$

for each $x \in A$. Similarly, f is said to be *bounded below* if its graph has no points under some line of equation $y = H$, i.e. if

$$f(x) \geq H,$$

for each $x \in A$. A function which is bounded both from below and above is said to be *bounded*.

Examples

Consider the following functions defined on $\mathbb{R}$ (their graphs are shown in the figure in the same order).

6.1. $x \mapsto 3x + 4$ is neither bounded above, nor below.

6.2. $x \mapsto x^2$ is bounded below (by $H = 0$), but not above.

6.3. $x \mapsto \dfrac{1}{1+x^2}$ is bounded, because $0 < \dfrac{1}{1+x^2} \leq 1$ for each $x \in \mathbb{R}$.

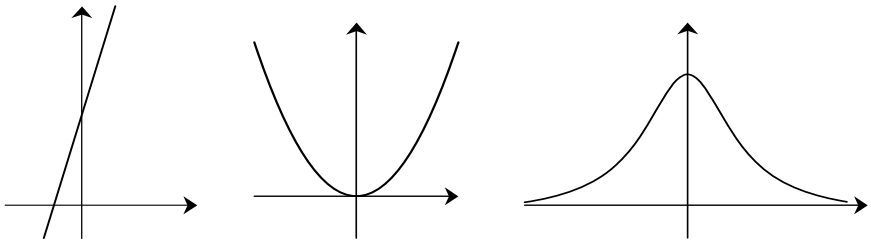


Figure 2.12. The three functions of Examples 6.1, 6.2, 6.3

2.6.2 Monotonic functions and sequences

Definition 6.1. *Consider $f : A \to \mathbb{R}$. If for each pair of points x_1 and x_2 in A*

$$x_1 < x_2 \qquad \textit{implies} \qquad f(x_1) \leq f(x_2) \tag{2.7}$$

then f is said to be (weakly) ***increasing*** *or not decreasing. If*

$$x_1 < x_2 \qquad \textit{implies} \qquad f(x_1) \geq f(x_2), \tag{2.8}$$

f is said to be (weakly) ***decreasing*** *or not increasing.*

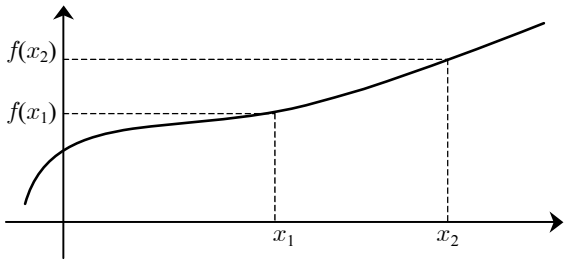


Figure 2.13. A strictly increasing function

If the strict inequalities are satisfied ($f(x_1) < f(x_2)$ and $f(x_1) > f(x_2)$ respectively), we say that f is ***strictly increasing*** *or* ***strictly decreasing.***

For example, the function $x \mapsto x^3$ is strictly increasing, and the function $x \mapsto k$ (constant) is both increasing and decreasing. Increasing and decreasing functions, in the weak or strict sense, are called *monotonic.*

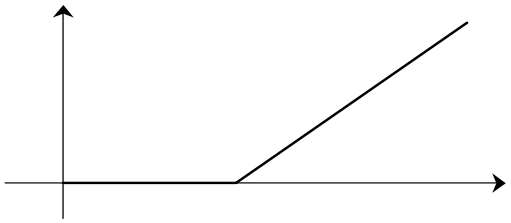


Figure 2.14. The function $\max(x - E, 0)$

The increasing function $f(x) = \max(x - E, 0)$, in Figure 14, is used in Finance to describe the final payoff of a *call option* versus the price x of the underlying activity.

To be clearer, we review the definitions of monotonic sequences.

Definition 6.2. *Let* $\{a_n\}$ be a *real value sequence. If for each* n *the inequality*

$$a_{n+1} \geq a_n$$

is satisfied, then $\{a_n\}$ *is said to be* (*weakly*) ***increasing*** *or not decreasing. If*

$$a_{n+1} \leq a_n,$$

then $\{a_n\}$ *is said to be* (*weakly*) ***decreasing*** *or not increasing.*

If the inequalities are strictly satisfied ($a_{n+1} > a_n$ *or* $a_{n+1} < a_n$), *we say that* $\{a_n\}$ *is* ***strictly increasing*** *or* ***strictly decreasing****, respectively.*

Example 6.4. The geometric sequence

$$a_n = q^n$$

assumes different behaviours according to the value of the ratio q.

If $q > 1$, it is strictly increasing, for example

$$1, 2, 4, 8, \ldots, 2^n, \ldots$$

If $q = 1$, it is constant and therefore both (weakly) increasing and decreasing.

If $0 < q < 1$, it is strictly decreasing, for example

$$1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots, \frac{1}{2^n} \ldots$$

If $q = 0$, it is constant again, while for $q < 0$ it is neither increasing nor decreasing.

2.6.3 Maximum and minimum values

We saw that some points on the graph of f correspond to a maximum or minimum height. Now we give precise definitions. Let, now, f be a function with domain A.

Definitions 6.3. *A real number M is called the (global)* ***maximum*** *of f in A and $x_0 \in A$ is called the (global)* ***maximum point***, *if, for each $x \in A$,*

$$M = f(x_0) \geq f(x). \tag{2.9}$$

Similarly, a real number m is called the (global) ***minimum*** *of f in A and $x_0 \in A$ is called the (global)* ***minimum point***, *if, for each $x \in A$,*

$$m = f(x_0) \leq f(x). \tag{2.10}$$

A maximum (or minimum) is said to be ***strict*** *if in (2.9) (or in (2.10)) the equal sign holds only for $x = x_0$.*

Maximum and minimum points are also called *extremum points* and maxima and minima are also called *extrema*. If maximum and minimum of f do exist, they are unique (why?). Maximum and minimum points may be many, and also infinitely many[11].

2.6.4 Convex and concave functions

Special subsets of the plane are called *convex.* For example, of the two polygons in Figure 15, the first is convex and the other is not. Circles, ellipses, half-planes are examples of convex sets.

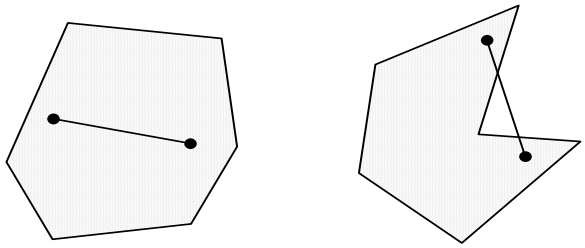


Figure 2.15. Convex and non-convex polygons

In general, a set of points in the plane is said to be *convex* if the segment connecting any pair of its points entirely lies in the set. The notion of convexity is easily transferred to functions, considering their *epigraph.* The *epigraph* of $f : A \to \mathbb{R}$ is the set consisting of all points (x, y) of the plane lying above the graph itself. In other words, the epigraph consists of all pairs (x, y), such that $x \in A$ and $y \geq f(x)$.

Definition 6.4. *f is said to be* ***convex*** *if its epigraph is convex.*

This is equivalent to requiring that each segment connecting two points on the graph of f lies entirely above or at least not under the graph of f, that is:

$f : (a, b) \to \mathbb{R}$ is convex if and only if for each pair $x_1, x_2 \in (a, b)$ and for each $t \in [0, 1]$,

$$f[tx_1 + (1-t)x_2] \leq tf(x_1) + (1-t)f(x_2).$$

[11] The reader may convince himself/herself with some examples.

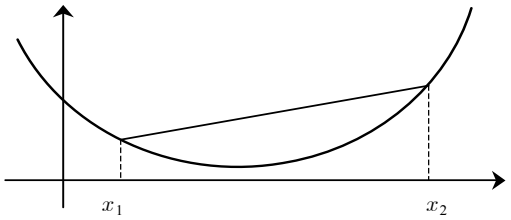


Figure 2.16. Graph of a strictly convex function

In fact, given two real numbers $x_1 < x_2$, the general point $\overline{x} \in [x_1, x_2]$ may be written as $\overline{x} = tx_1 + (1-t)\,x_2$, with $t \in [0,1]$, and the y-coordinate of the general point which lies on the straight line connecting points $(x_1, f(x_1))$ and $(x_2, f(x_2))$ is $\overline{y} = tf(x_1) + (1-t)\,f(x_2)$.

We note that the graph of a convex function may include line segments (straight parts of graph). If this does not happen, we say that the function is *strictly convex.*

Typical strictly convex functions are parabolae $f(x) = ax^2$, if a is positive.

Definition 6.5. *f is* ***concave*** *(strictly concave) if $-f$ is convex (strictly convex).*

Functions $x \mapsto ax^2$ are strictly concave, if a is negative.

It follows from the definition that the domain of convex and concave functions can only be an *interval* (which is the only kind of convex set in $\mathbb{R}$). The notions of convexity and concavity allow us to find special points in the graph of a function f, called *points of inflection* . Intuitively we are dealing with points where the graph "changes its concavity".

Definition 6.6. *A point x_0 in the domain of f is said to be a* ***point of inflection*** *if it is possible to find an interval $(x_0 - \delta, x_0]$, on the left of x_0, and an interval $[x_0, x_0 + \delta)$, on the right of x_0, where f turns out to be respectively convex and concave (or concave and convex).*

A typical case is the cubic function $f(x) = x^3$, which has an inflection point at the origin (see Figure 17).

2.6.5 Local properties

Sometimes we might be interested in properties which are satisfied by a function near to a point x_0 of its domain. We then talk about *local* properties, in contrast with properties satisfied in its whole domain (the so-called *global* properties).

For example, the function $f(x) = x^2$, which is not (globally) increasing, turns out to be increasing "near" to $x = 1$. We say that it is *locally increasing.*

The point $x = 2$ is not a global maximum point for the function in Figure 18. However, if we consider the function only in an interval of the kind $(2 - \delta, 2 + \delta)$, with $\delta > 0$ small enough, the point $x = 2$ turns out to be a maximum point. Then we

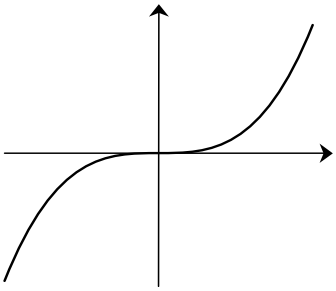


Figure 2.17. Inflection point at the origin for $f(x) = x^3$

say that the point is a *local maximum* point. This notion is particularly important and deserves a precise formulation.

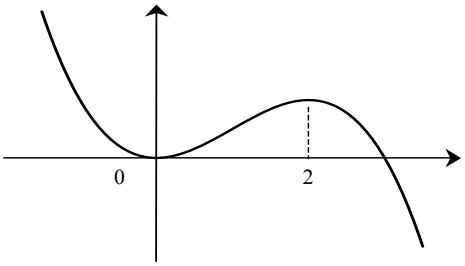


Figure 2.18. Local minimum and maximum

Definition 6.7. *Let x_0 be a point in the domain of f. If there is a neighbourhood $(x_0 - \delta, x_0 + \delta)$, with $\delta > 0$, such that, for every x in such a neighbourhood and in the domain of f, we have*

$$f(x_0) \geq f(x), \tag{2.11}$$

then x_0 is said to be a ***local maximum point****. If*

$$f(x_0) \leq f(x), \tag{2.12}$$

then x_0 is said to be a ***local minimum point****. If in* (2.11) *and* (2.12) *the equal sign is attained only for $x = x_0$, then the maximum and minimum are said to be* ***strict****.*

It should be clear that global properties, such as being positive, increasing, convex, are also local. In general the converse is not true, in the sense that a property, which is locally satisfied, is not necessarily globally valid.

In particular: *each global maximum or minimum point is also a local maximum or minimum point.*

2.7 Power functions

The functions $f(x) = kx,\ f(x) = kx^2,\ f(x) = k/x$ we met in sections 2 and 3 are all examples of *power functions*, as they can be written in the form

$$\boxed{f\,(x) = kx^{\alpha}} \qquad \alpha \neq 0.$$

Generally speaking, they are defined only for $x > 0$. If $\alpha > 0$ they are also defined at $x = 0$, where their value is 0. Some "privileged member" of the family may be defined on the whole set $\mathbb{R}$, as shown by the examples $y = x^2$ and $y = x^3$. Functions $y = 1/x = x^{-1}$ and $y = 1/x^2 = x^{-2}$ are defined on $\mathbb{R} \setminus \{0\}$, while the function $y = \sqrt{x} = x^{1/2}$ is defined on $[0, +\infty)$.

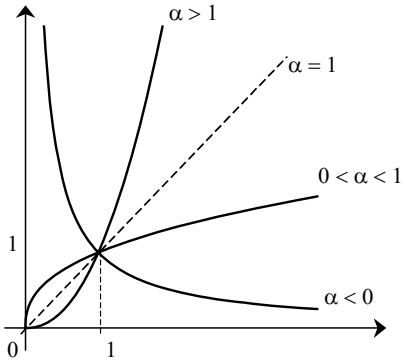


Figure 2.19. Power functions

Figure 19 illustrates graphs of power functions, with different values of α, in the case $k = 1$. The graphs show

— *strictly increasing* functions for $\alpha > 0$, *strictly decreasing* functions for $\alpha < 0$,

— *strictly convex* functions for $\alpha < 0$, $\alpha > 1$ and *strictly concave* functions for $0 < \alpha < 1$.

As x grows larger and larger, it can be seen that, if $\alpha > 0$, the power x^{α} assumes larger and larger values, while if $\alpha < 0$, it approaches 0. Further on we will express these facts saying that, for x tending to $+\infty$, the power x^{α} tends to $+\infty$ in the first case and to 0 in the second case.

When x takes values which are closer and closer to zero, powers with negative exponent attain larger and larger values. We will say that when x tends to zero the power x^{α} tends to $+\infty$.

• *Inventory management.* Further in the book, we will see that a firm looking for the optimal amount Q of the raw material to purchase should order, each time,

$$Q = \sqrt{\frac{2Sg}{m}}$$

where S is the quantity of raw material needed for one year, g is the fixed cost for managing one order and m is the cost for holding in stock one unit of material over one year. If we consider Q as a function of the total requirements or a function of the cost of one order, we have a power function with $\alpha = 1/2$; if we consider Q as a function of m we have an example where $\alpha = -1/2$.

Even and odd symmetry

The graphs of even exponent powers, like $y = x^2$ or $y = x^4$, are symmetric with respect to the y-axis. If the exponent is odd, like $y = x^3$, the graphs are symmetric with respect to the origin. Obviously, not only power functions show these sorts of symmetry. In general we say that a function f is **even** if its graph is symmetric with respect to the y-axis, that is points with opposite x-coordinates have the same y-coordinate:

$$f(-x) = f(x)$$

On the other hand, if the graph is symmetric with respect to the origin, we say that f is **odd**; in this case, points with opposite x-coordinates have opposite y-coordinates:

$$f(-x) = -f(x)$$

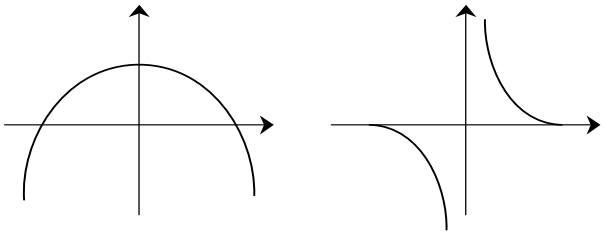


Figure 2.20. Graphs of even and odd functions

Invertibility of power functions

The function $y = x$ is the inverse function of itself, as is the function $y = 1/x$.

The function $y = x^3$ is invertible: its inverse function is $y = \sqrt[3]{x}$.

The function $y = x^2$, defined on $\mathbb{R}$, is not invertible. Each horizontal line lying in the upper half-plane intersects the graph (a parabola) in two points, symmetric with respect to the y-axis. If we forget the left "semi-parabola" and consider the function $y = x^2$ as defined only on $[0, +\infty)$, then we get an invertible function. Its inverse function is the *arithmetic square root*: $y = \sqrt{x}$.

Power functions $y = x^\alpha$ $(\alpha \neq 0)$, defined only for $x > 0$, are invertible and their inverses are power functions again, but with $1/\alpha$ as an exponent. More precisely: the inverse function of $f(x) = x^\alpha$ is $f^{-1}(x) = x^{1/\alpha}$.

Warning! Do not confuse $f^{-1}(x)$ with $1/f\,(x)$! If a negative exponent is preferred, $1/f\,(x)$ can be written as $[f\,(x)]^{-1}$.

If $f(x) = x^3$, we get $f^{-1}(x) = \sqrt[3]{x}$, while $[f\,(x)]^{-1} = \dfrac{1}{f\,(x)} = \dfrac{1}{x^3}$.

2.8 Exponential, logarithmic, trigonometric functions

Together with power functions, exponential, logarithmic and trigonometric functions are bricks used for building a considerable amount of mathematical models in many applications. Let us review their main properties, possibly equipped with a computer and a graphic software.

2.8.1 Exponential function

Often in economic models the decay in time of instruments (like machines, plants,...), the productivity, the growth of a market share, the continuous time dynamics of an invested capital are described by means of exponential functions. The function

$$f(x) = a^x,$$

where a is a positive real number different[12] from 1, is called the *exponential function* with base a.

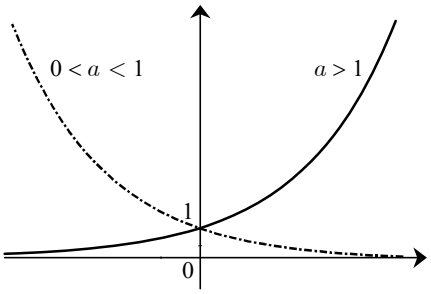


Figure 2.21. Exponential functions

Exponential functions are *positive* for each value of a and their graph passes through the point $(0, 1)$. A glance at the graphs reveals that they are *strictly increasing* if $a > 1$, *strictly decreasing* if $0 < a < 1$. For each a they are *strictly convex* functions.

If $a > 1$, for larger and larger values of the exponent x we get larger and larger values for the exponential a^x, while if $0 < a < 1$ we get values which approach 0. Speeding things up again, we will say: as x *tends to* $+\infty$, a^x *tends to* $+\infty$ in the first case, to 0 in the second case.

[12] The case $a = 1$ is not interesting.

We note, by the way, that since

$$\left(\frac{1}{a}\right)^x = a^{-x},$$

two exponential functions whose bases are reciprocal have graphs which are symmetric with respect to the y-axis.

2.8.2 Logarithmic functions

Each exponential function is invertible, as the graphs clearly show. The inverse function of $f(x) = a^x$ $(a > 0,\ a \neq 1)$ is called the *logarithmic function* and it is denoted by the symbol

$$f^{-1}(x) = \log_a x.$$

From what we have already seen in the first chapter, the domain of a logarithmic function is the interval $(0, +\infty)$. The graphs of logarithmic functions are obtained from the graphs of the corresponding exponential functions, by means of a reflection with respect to the bisector of the first and third quadrants.

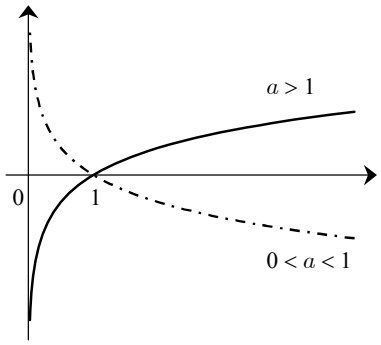


Figure 2.22. Logarithmic functions

It is sufficient to look at the graph to convince oneself that:

— if $a > 1$, the logarithmic function is *strictly increasing and concave;* positive for $x > 1$, negative for $0 < x < 1$.

— if $0 < a < 1$, it is *strictly decreasing and convex;* positive for $0 < x < 1$, negative for $x > 1$.

Note that we have $\log_a 1 = 0$, for every base a.

Let us have a look at what happens near to the endpoints of the domain. For values of x approaching zero, $\log_a x$ assumes larger and larger absolute values, negative if $a > 1$, positive if $0 < a < 1$. We say that, as x tends to zero, $\log_a x$ tends to $-\infty$ or $+\infty$.

If the independent variable assumes larger and larger values, the logarithm also assumes larger and larger absolute values, positive if $a > 1$, negative if $0 < a < 1$. We say that, as x tends to $+\infty$, $\log_a x$ tends to $-\infty$ or $+\infty$.

There is a privileged base for exponential and logarithmic functions: it is the (irrational) *Napier's number*[13] $e \simeq 2.71828\ldots$, which will be defined more formally in chapter 3. In what follows, in order to refer to logarithms with base e, we will use the symbol ln (instead of $\log_e$) which stands for *natural logarithm*. The exponential function with base e is also indicated by $\exp(x)$.

2.8.3 Trigonometric functions

Trigonometric functions are used in models describing periodic phenomena, like the motion of planets, (mechanic or electromagnetic) wave propagation, the course of certain flu outbreaks of a seasonal nature.

In economic applications they have been used to describe typically cyclical phenomena, like the growth of an economic system, where periods of development alternate with recessions. A historical example (1938) is represented by P.A. Samuelson's *accelerator model*.

The functions $y = \sin x$ (sine of x) and $y = \cos x$ (cosine of x) may be introduced in the following way. Consider the unit circle having its centre at the origin, as in figure 23. Let A be the point with coordinates $(1, 0)$ and let x be the length of the arc AB, with x positive if we are moving from A to B counter-clock-wise, negative if moving clock-wise; since the radius of the disc is 1, x is also the *measure in radians* of the angle[14] AOB.

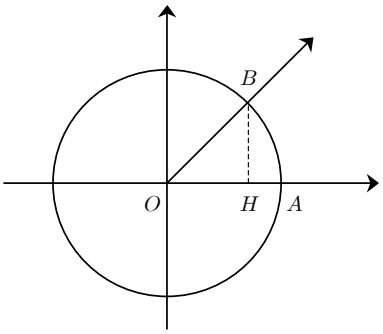


Figure 2.23.

The function associating with x the vertical coordinate of B is called the *sine of* x and it is denoted by the symbol $\sin x$; the function associating with x the horizontal coordinate of B is called the *cosine of* x and it is denoted by the symbol $\cos x$. Obviously, adding to x (positive or negative) multiples of 2π, the coordinates of the point B do not change: after a full circle $(x + 2\pi)$ the point B ends up in the starting

[13] John Napier (1550-1617), Scottish mathematician. Dealing with logarithms, he had crucial intuitions about the continuity of curves - a subject which will be considered later.

[14] An angle of one radian corresponds to a straight angle divided by π.

position. Thus we have, for each real x,

$$\sin(x + 2\pi) = \sin x, \qquad \cos(x + 2\pi) = \cos x$$

and we say that the functions *sine and cosine are periodic with period* 2π. Their graphs are drawn in figures 24 and 25.

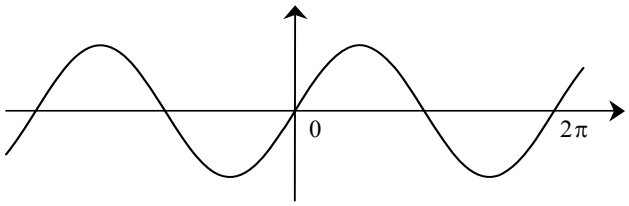


Figure 2.24. Graph of $\sin x$

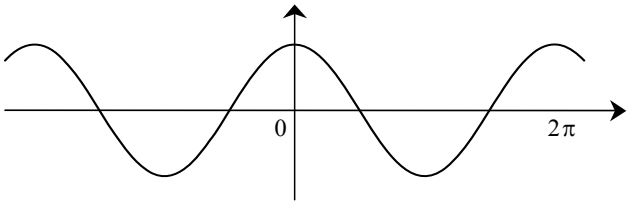


Figure 2.25. Graph of $\cos x$

Generally speaking, a function $f : \mathbb{R} \to \mathbb{R}$ is said to be *periodic* with period T, if T is the smallest positive number such that, for every x in the domain of f,

$$f(x + T) = f(x).$$

Let us consider some simple properties of sine and cosine.

Fundamental identity. Directly from the definition, we get:

$$(\cos x)^2 + (\sin x)^2 = 1.$$

Symmetry. The function sine is *odd*, the function cosine is *even*:

$$\sin(-x) = -\sin x, \qquad \cos(-x) = \cos x.$$

Boundedness. The functions sine and cosine are *bounded*. The value 1 is the *maximum* for both, the value -1 is the *minimum* for both. Because of the periodicity, the maximum and minimum points are infinitely many and are, respectively,

$$x = \frac{\pi}{2} + 2k\pi \quad \text{and} \quad x = -\frac{\pi}{2} + 2k\pi, \qquad k \in \mathbb{Z}$$

for sine, and

$$x = 2k\pi \quad \text{and} \quad x = -\pi + 2k\pi, \qquad k \in \mathbb{Z}$$

for cosine.

Shift formulae. The cosine graph is the sine graph shifted $\pi/2$ to the left and, vice versa, the sine graph is the cosine graph shifted $\pi/2$ to the right. In formulae this can be translated as

$$\cos x = \sin\left(x + \frac{\pi}{2}\right), \qquad \sin x = \cos\left(x - \frac{\pi}{2}\right).$$

Sum and difference formulae:

$$\begin{aligned} \sin(\alpha \pm \beta) &= \sin\alpha\cos\beta \pm \cos\alpha\sin\beta \\ \cos(\alpha \pm \beta) &= \cos\alpha\cos\beta \mp \sin\alpha\sin\beta. \end{aligned}$$

In particular,

$$\sin 2\alpha = 2\sin\alpha\cos\alpha, \qquad \cos 2\alpha = (\cos\alpha)^2 - (\sin\alpha)^2.$$

Another trigonometric function, the *tangent* $x \mapsto \tan x$, may be defined as the ratio between sine and cosine:

$$\tan x = \frac{\sin x}{\cos x}$$

or as the ratio between the vertical and horizontal coordinates of point B (Figure 23). A great part of its importance is due to the fact that it represents the *slope* of the straight line connecting points O and B. The tangent function is *odd,* it is defined for values of x such that the cosine is different from zero, that is

$$x \neq \frac{\pi}{2} + k\pi, \qquad k \in \mathbb{Z}$$

and it is *periodic with period* π. Its graph is drawn in Figure 26.

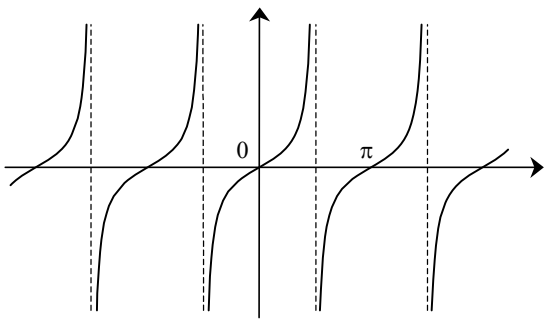


Figure 2.26. Graph of $\tan x$

Invertibility of trigonometric functions. Given their periodicity, if trigonometric functions are considered over their whole domain, they are not invertible. In order to

obtain invertible functions we need to restrict the domain. We limit ourselves to the case of the tangent and we restrict our considerations to the interval $(-\pi/2, \pi/2)$; then the corresponding part of the graph is associated with an invertible function. Its inverse function is called the *inverse tangent*, it is defined on $\mathbb{R}$ and takes its values in $(-\pi/2, \pi/2)$. Its graph is shown in Figure 27.

The function $x \mapsto \tan^{-1} x$ is *odd, increasing* and *bounded*.

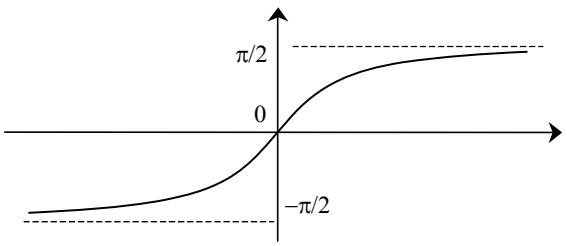


Figure 2.27. Graph of $\tan^{-1} x$

2.9 Geometric transformations

In this section we shall consider simple examples of composite functions, interpreting their action in terms of geometric transformations in the plane. The reader is invited to verify our results with the help of a computer and a graphic software.

• The function $x \mapsto x + k$ is a *shift*.

It is equivalent to "moving the origin": the coordinate system changes without altering the unit of measure. For example, given a body temperature t in Celsius degrees, the absolute temperature T (Kelvin degrees) is obtained by the shift $t \mapsto T = t + 273.14$.

A function f may be composed with a shift, getting the functions

$$x \mapsto f(x) + k \qquad \text{and} \qquad x \mapsto f(x + k).$$

The graph of the first function is obtained via a shift of the graph of f by k units upwards if k is positive, by $-k$ units downwards if k is negative; the graph of the second function by moving the graph of f left by k units if k is positive, right by $-k$ units if k is negative.

• *Homographic function.* The graph of the function

$$f(x) = \frac{ax + b}{x + c}, \quad a, b, c \in \mathbb{R},\ b \neq ac$$

is an equilateral hyperbola. The *asymptotes* are the two straight lines

$$x = -c \qquad \text{and} \qquad y = a$$

parallel to the vertical and horizontal axes respectively. Its graph is obtained by shifting the hyperbola $y = k/x$ ($k = b - ac$) in the direction of the axes. Let us draw, for instance, the graph of the function

$$f(x) = \frac{2x}{x-1}.$$

It can be written as

$$f(x) = 2 + \frac{2}{x-1}.$$

The graph of f is obtained by moving the graph of the function $y = 2/x$ up by two units and right by one unit.

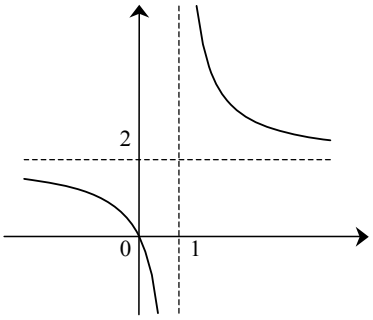


Figure 2.28. Graph of $f(x) = \dfrac{2x}{x-1}$

Let $C = C(q)$ be the total production cost, as a function of the produced amount q. The function $C_{\mathrm{M}}(q) = C(q)/q$ represents the *unitary average cost.* If C is affine linear, then C_{M} is a homographic function.

• In *local studies*, that is when we are looking for information about the behaviour of a function f "near" to a point x_0, it may turn out to be convenient to make x_0 coincide with the zero point, and to make the value assumed by the function be zero. In this case, instead of $f(x)$, we consider the function g in the new variable h defined by

$$g(h) = f(x_0 + h) - f(x_0).$$

This is equivalent, from a graphical point of view, to shifting the axes, so that the new origin coincides with the point $(x_0, f(x_0))$. Such transformations are used when we are interested in the analysis of deviations of a given quantity (f in this case) with respect to a reference situation (represented here by the point $(x_0, f(x_0))$.

• The linear function $x \mapsto kx$ may be read, if k is a positive number, as a "change of the unit of measure".

For example, in order to calculate the price of an item in Euro when its cost is x dollars[15], we use the function $x \mapsto 0.8x$; to calculate in meters a distance x which is expressed in miles we use the function $x \mapsto 1600x$.

[15] Taking the exchange rate as 0.8 euro for one dollar.

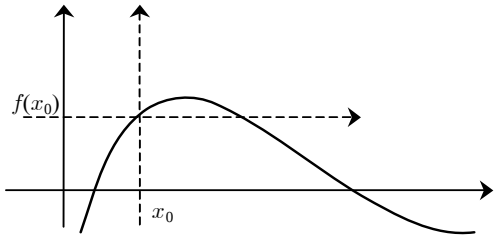


Figure 2.29.

A change in the unit of measure on the real axis makes the graph shrink or stretch. If k is negative, there is also a change of orientation.

A function f may be composed with a linear function, giving the functions

$$x \mapsto kf(x) \qquad \text{and} \qquad x \mapsto f(kx),$$

corresponding to changes in the unit of measure on the y-axis and x-axis.

Considering the function $x \mapsto kf(x)$, if $k > 1$ the vertical coordinates on the graph of f are multiplied by the factor k; so that the graph of kf is obtained by "stretching" the graph of f in the vertical direction, that is expanding upwards all positive vertical coordinates and downwards the negative ones; if $0 < k < 1$ the graph is shrunk, again in the vertical direction.

If k is negative a reflection with respect to the x-axis is also associated. In particular, if $k = -1$ vertical coordinates change their sign, so that the graph of $-f$ turns out to be symmetric, with respect to the x-axis, with the graph of f.

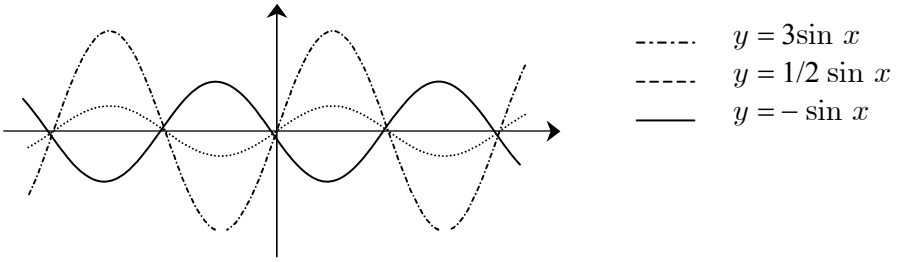


Figure 2.30.

In Figure 30 we illustrate graphs of the functions $y = 3\sin x$, $y = \dfrac{1}{2}\sin x$ and $y = -\sin x$, obtained from the graph of $y = \sin x$.

If $k > 1$, the graph of $x \mapsto f(kx)$ is obtained by shrinking the graph of f in the horizontal direction, while if $0 < k < 1$ the graph is stretched, again in the horizontal direction. To see this, it is enough to think that if the domain of f is the interval $[6, 15]$, kx varies from 6 to 15 and, consequently, x varies from $6/k$ and $15/k$. If now $k > 1$, for example $k = 3$, the interval $[6, 15]$ shrinks its length by a factor 3 and

moves towards the origin turning into the interval $[2, 5]$. On the contrary, if $k = 1/2$, the new interval doubles its length and moves away from the origin turning into the interval $[12, 30]$.

If k is negative a reflection with respect to the y-axis is also associated. In particular, if $k = -1$, horizontal coordinates change their sign, so that the graph of $y = f(-x)$ is symmetric, with respect to the y-axis, with the graph of f.

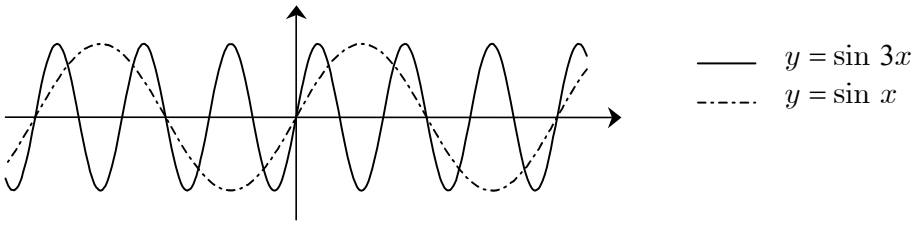


Figure 2.31.

In Figure 31 we show graphs of the functions $y = \sin x$ and $y = \sin 3x$.

• *The absolute value function.* Composing a function f with the function $x \mapsto |x|$, whose graph is drawn in Figure 32, we get the functions

$$x \mapsto |f(x)| \qquad \text{and} \qquad x \mapsto f(|x|).$$

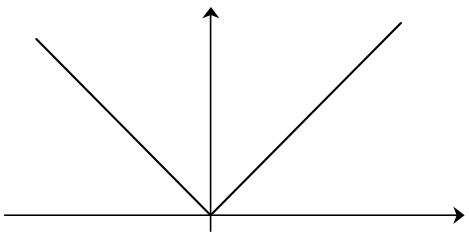


Figure 2.32. Graph of the function $y = |x|$

Remembering that

$$|f(x)| = \begin{cases} f(x) & \text{if } f(x) \geq 0 \\ -f(x) & \text{if } f(x) < 0, \end{cases}$$

we deduce that the graph of $y = |f(x)|$ is obtained from the graph of f by "reflecting" above the horizontal axis the underlying part of the graph.

On the other hand, the graph of $y = f(|x|)$ is obtained from the graph of f by substituting the part of graph lying on the left of the vertical axis with a curve which is the symmetric with respect to the y-axis of the part of the graph lying on the right:

$$f(|x|) = \begin{cases} f(x) & \text{if } x \geq 0 \\ f(-x) & \text{if } x < 0. \end{cases}$$

Whatever initial function f is considered, the function $y = f(|x|)$ is even.

Figure 33 shows the graphs of the functions $y = \ln x$, $y = |\ln x|$ and $y = \ln |x|$.

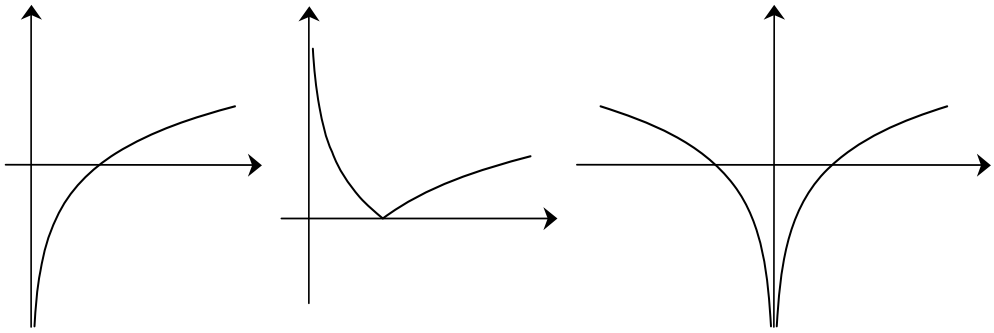


Figure 2.33. Graphs of $y = \ln x$, $y = |\ln x|$ and $y = \ln |x|$

2.10 Exercises

2.1. Let A and B be the natural domains of the functions

$$f(x) = \ln\left(x^2 - x\right), \qquad g(x) = \ln x + \ln(x-1).$$

Establish which of the following relations is true:

$$A \subset B, \qquad A = B, \qquad B \subset A.$$

2.2. Determine the natural domains of the functions

$$f(x) = \sqrt[6]{\frac{x^2 - 3x + 2}{3 - x}}, \qquad g(x) = \sqrt[7]{\frac{x^2 - 3x + 2}{3 - x}}.$$

2.3. Let the function

$$D(x) = \begin{cases} -\dfrac{x}{5} + 70 & 100 \leq x < 200 \\ -\dfrac{x}{20} + 40 & 200 \leq x \leq 500 \end{cases}$$

represent the demanded amount of some goods as a function of the price x. Construct the function $r(x) = xD(x)$ (revenue) and draw its graph. For what price do we get the maximum revenue?

2.4. Draw the graph of the demand function

$$q = q(p) = 2000e^{-p/10} \quad (p > 0).$$

Check that it is invertible and deduce the price p as a function of the produced quantity $p = p(q)$.

2.5. (⇛ **Chapter 11**) The accumulation factor in anticipated simple interest is

$$f(t) = \frac{1}{1-dt}.$$

Draw the graph of the function $f(t)$ on the interval $[0, 1/d)$, corresponding to the values $d_1 = 5\%$ and $d_2 = 10\%$.

2.6. (⇛ **Chapter 11**) The DCF of a financial operation (a 9 Euro investment today yielding 10 Euro in two years) is

$$G_1(x) = -9 + \frac{10}{(1+x)^2}.$$

Draw the graph of the function G_1.

Looking at it "from the other side" (we get 9 Euro today and we will pay 10 Euro in two years), the DCF is

$$G_2(x) = 9 - \frac{10}{(1+x)^2}.$$

How can we obtain the graph of G_2 from the graph of G_1?

2.7. Let $f(x) = 1 + \dfrac{1}{x}$, $x > 0$. Write the inverse function of f and the composite function $f \circ f$.

2.8. Solve, algebraically and graphically, the following inequalities

$$(a)\ \sqrt{x+3} > 1 + x, \qquad (b)\ \sqrt{x+5} < x - 1.$$

2.9. Check that the function $f(x) = \sin(2\pi x)$ is periodic with period $p = 1$. How should we modify f in order to have 3 complete oscillations in the interval $[0, 1]$, instead of one?

2.10. Prove that, if f and g are increasing, then $f + g$ and $f \circ g$ are also increasing.

What can we say in the case of decreasing functions or in the case of one increasing function and one decreasing? What can we say of their product and their quotient?

2.11. Which of the following functions are monotonic? Which are even or odd?

$$f(x) = \ln(1-x), \quad g(x) = x^3 + x, \quad h(x) = e^{-x} - x, \quad k(x) = \frac{\sin x}{x}.$$

2.12. A photocopy service may be equipped with a normal machine model, having a low fixed cost (f) and a high variable unitary cost (V). As an alternative, there is the advanced model with a high fixed cost (F) and a lower variable unitary cost

(v). Letting x be the number of copies to be made, from an economic point of view when does the advanced model turn out to be profitable? Give a graphic description of the situation.

2.13. The demand for goods varies as a function of price according to the law $q(p) = 100 - p$, while the production cost varies as a function of the quantity produced according to the law $C(q) = 400 + 2q$. Draw the graphs of functions $r(p)$ and $r(q)$ (revenue as a function of price and quantity respectively) and $\pi(p)$ and $\pi(q)$ (profit as a function of price and quantity). Find the values of p and q for which the profit attains its maximum.

2.14. In some diffusion models for a product in a market, it is supposed that after x years the achieved market share is $q(x) = 1 - e^{-ax}$, where a is a positive number and $x \geq 0$. Starting from the graph of the exponential function, construct the graph of the function q for $x \geq 0$. Is it true that if x is sufficiently large the market share becomes as close to 100% as we wish?

2.15. The sales of a firm are exponentially increasing with time as described by the model $v(x) = he^{ax}$, with $a, h > 0$. Deduce the expression of the time x needed to make the sales level attain an amount V greater or equal to h. Give a graphic description of x as a function of V.

2.16. *The "zoom" effect.* The graph of the function $y = kf\left(\frac{x}{k}\right)$ is a magnification to scale k (> 1) of the graph of f. Using a computer, draw the graph of the function $f(x) = \sin x$, $x \in [-\pi, \pi]$. Then deduce the magnifications of the graph to scales 2, 3, 5.

3

Limits

The ideas and techniques of Calculus, which are the subject of our next chapters, are part of the essential tools of modern science. And Calculus rests on the notion of limit, a true milestone in the history of scientific thought, which we develop in this chapter along the following lines.

• The definition of *limit* is first introduced for *sequences*, and sequences are classified as *convergent, divergent* or *irregular*. We then define *infinities* and *infinitesimals.*

• We extend the definition of limit from sequences to functions. In particular,
— we define *right-hand limits*, *left-hand limits*, *bilateral limits*;
— we establish the existence of limits for monotonic sequences and functions, and apply this result to limits of elementary functions.

• We then arrive at the important definition of Napier's number e.

• We explore the behaviour of the limit operation with respect to algebraic operations and to comparisons among functions. We give appropriate theorems on:
— the limit of a sum, a product and a quotient;
— comparison and permanence of sign;
— the limit of a composite function.

• We finally deal with the problem of comparing different speeds of convergence and divergence. This naturally leads us to introduce some symbols ("asymptotic" and "little-o") which prove to be useful when coping with this question, and to establish a hierarchy among infinities and infinitesimals.

3.1 Limits of sequences

3.1.1 Asymptotic properties of a sequence

The term *sequence* is intuitively associated with an infinite list of objects. The question which should arise spontaneously when looking at a sequence is: "What happens if we proceed further and further in the list?". The answer is provided by the notion of limit.

We start with a preliminary definition. What happens "when we proceed further and further" depends on some properties which are satisfied by all the terms of a sequence $\{a_n\}$ when n is "large enough". This is what we really find interesting: not the properties which are satisfied by all terms a_n, but the properties which are satisfied by the terms corresponding to the values of n which start from a certain index onwards.

Definition 1.1. *The* ***asymptotic properties of a sequence*** $\{a_n\}$ *are those properties which are satisfied by its terms at least from a certain index* n_0 *onwards, that is for all values of the index* n *which are large enough.*

For example, all terms of the sequence $a_n = \dfrac{n-3}{n+1}$ are positive, if n is large enough; the first three terms are indeed negative, and the fourth one is null, but for $n > 3$ all terms are positive.

An equivalent definition of an asymptotic property of a sequence is a property which is satisfied by all the terms of the sequence *except for, at most, a finite number of them.*

3.1.2 Convergent sequences

The sequence $\{a_n\}$ defined by

$$a_n = \frac{n-1}{n} = 1 - \frac{1}{n}$$

never attains the value 1; but, when n is large enough, the term a_n gets as near to 1 as we want, because when n increases the term $1/n$ becomes arbitrarily small. As a consequence, a_n gets "indefinitely" near to 1. We can say this more precisely: given any number $\varepsilon > 0$,

$$\text{the } \textit{distance} \text{ of } a_n \text{ from 1 is less than } \varepsilon, \text{ if } n \text{ is large enough} \tag{3.1}$$

This is indeed equivalent to requiring that the condition

$$|a_n - 1| = \frac{1}{n} < \varepsilon \tag{3.2}$$

holds (at least) if n is large enough, for every $\varepsilon > 0$. Formula (3.2) is satisfied for all indices n such that

$$n > \frac{1}{\varepsilon}$$

that is, when n is large enough. We shall say that the sequence

$$a_n = \frac{n-1}{n} \quad \textit{converges to } 1.$$

Definition 1.2. *A sequence $\{a_n\}$ is called* ***convergent*** *to the real number A if, for any $\varepsilon > 0$, the distance of a_n from A is less than ε when n is large enough, that is if*

$$|a_n - A| < \varepsilon$$

when n is large enough.

This situation is usually described by writing

$$\boxed{\lim_{n\to+\infty} a_n = A} \qquad \text{or} \qquad \boxed{a_n \to A \text{ as } n \to +\infty} \tag{3.3}$$

which read "the limit of a_n as n tends to plus infinity is equal to A", or "a_n tends to (or converges to) A as n tends to plus infinity" respectively[1].

This does not mean at all that a_n needs to reach the value A, but simply that it gets indefinitely near to it. Naturally, if for a given sequence $\{a_n\}$ the equality $a_n = A$ is true whenever n is large enough, it follows that condition (3.3) holds.

If $a_n \to A$, and $a_n \geq A$ when n is large enough, we say that a_n converges to A by excess and write

$$a_n \to A^+ \text{ as } n \to +\infty \quad \text{or} \quad \lim_{n\to+\infty} a_n = A^+.$$

Analogously, if $a_n \leq A$ when n is large enough, we say that a_n converges to A by defect and use the symbol A^-.

If we refer to the geometrical representation of a sequence, saying that a sequence a_n converges to the limit A as $n \to +\infty$ means that, however we might choose a horizontal strip, centred at height A and with an arbitrarily small amplitude, at least from a certain term onwards all points of the graph with coordinates (n, a_n) are inside the strip.

A sequence which converges to zero is said to be an infinitesimal sequence, or simply an **infinitesimal**. Examples of *infinitesimal* sequences are

$$\left\{\frac{1}{n}\right\}, \quad \left\{\frac{1}{2^n}\right\}, \quad \left\{\frac{1}{n^2}\right\}.$$

Let us check, for example, that $\lim_{n\to+\infty} \dfrac{1}{n^2} = 0$. This is equivalent to requiring that the condition

$$\left|\frac{1}{n^2} - 0\right| = \frac{1}{n^2} < \varepsilon \tag{3.4}$$

[1] We can also write $\lim a_n = A$ or $a_n \to A$, and leave the clause that $n \to +\infty$ implicit.

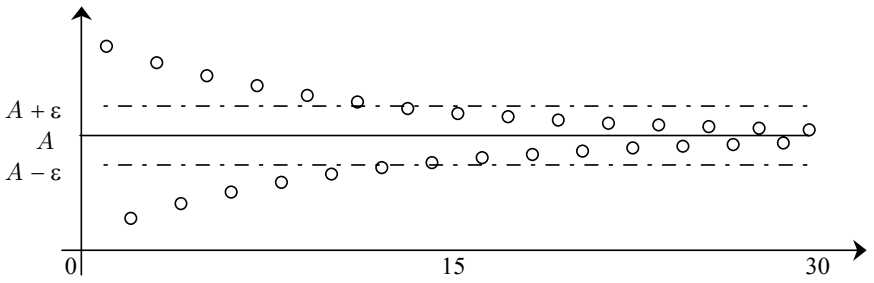


Figure 3.1. Graph of a convergent sequence

holds (at least) if n is large enough, for every $\varepsilon > 0$. The condition (3.4) is satisfied by all indices n such that

$$n > \sqrt{\frac{1}{\varepsilon}}$$

that is, when n is large enough.

Remark 1.1. The three following statements are equivalent, i.e. each one of them implies the other two.

$$a_n \to A; \quad a_n - A \to 0; \quad |a_n - A| \to 0.$$

3.1.3 Divergent sequences

Let us consider

$$a_n = 3n - 2.$$

However we might choose a number M, and require that it be a sort of "ceiling" for the values of the sequence $\{a_n\}$, M will be exceeded by a_n when n is large enough. In fact, the condition $a_n > M$ is equivalent to $3n - 2 > M$, and it is satisfied by all integers n such that

$$n > \frac{M-2}{3}.$$

We will say that this sequence *diverges to* $+\infty$ when $n \to +\infty$. The graph of such a sequence will cross above any horizontal straight line of equation $y = M$, when n is large enough, however large we might choose the number M.

The opposite sequence, that is the sequence $b_n = -3n + 2$, satisfies the condition $b_n < M$ when n is large enough, however we might choose M (and here the case of M negative is the interesting one): we will say that this sequence *diverges to* $-\infty$. For example, the sequence

$$a_n = -\sqrt{n}$$

diverges to $-\infty$. In fact, the condition

$$a_n = -\sqrt{n} < M$$

is trivially true if $M > 0$, otherwise it holds only from a certain term onwards; that is, when

$$n > M^2,$$

and this means that in any case it holds (at least) if n is large enough. The graph of such a sequence will cross under any horizontal straight line of equation $y = M$, when n is large enough, however we might choose the number M.

Definition 1.3. *A sequence $\{a_n\}$ is called* ***divergent*** *to $+\infty$ ($-\infty$) if, for any real number M, we have*

$$a_n > M \qquad (a_n < M)$$

when n is large enough. In this case, we write

$$\boxed{\lim_{n\to+\infty} a_n = +\infty} \qquad \text{or} \qquad \boxed{a_n \to +\infty, \text{ as } n \to +\infty}$$

($\lim_{n\to+\infty} a_n = -\infty$ or $a_n \to -\infty$, as $n \to +\infty$ respectively), which read "the limit of a_n as n tends to plus infinity is equal to plus infinity (minus infinity)", or "a_n tends to or diverges to plus infinity (minus infinity) as n tends to plus infinity" respectively.

A divergent sequence is also called an infinite sequence, or simply an **infinity**. Examples of *infinite* sequences are

$$\{n\}, \quad \left\{n^2\right\}, \quad \left\{2^n\right\}, \quad \left\{-n^3\right\}.$$

If the sequence $\{a_n\}$ is an infinity, then the sequence $\{b_n\}$ whose general term is

$$b_n = \frac{1}{a_n}$$

is called the *reciprocal* sequence of $\{a_n\}$ and is an *infinitesimal*. For example, the sequence

$$b_n = \frac{1}{3n-2}$$

is infinitesimal.

- (⇛ **Chapter 11**) *Accumulation factors and discount factors.* The sequences

$$a_n = 1 + in \quad \text{and} \quad b_n = (1+i)^n \quad (i > 0)$$

correspond to the accumulation factors with simple interest and compound interest, and they are infinite sequences. The related discount factors

$$c_n = \frac{1}{1+in} \qquad \text{and} \qquad d_n = \frac{1}{(1+i)^n} \qquad (i > 0)$$

are infinitesimals.

3.1.4 Irregular sequences

There exist sequences which are neither convergent nor divergent. For example, $a_n = (-1)^n$ oscillates between -1 and 1. The distance of its terms from the number 1 (or from the number -1) cannot be less than an arbitrary ε for all values of n when n is large enough, if ε is chosen to be < 2. Another example is $b_n = (-1)^n n^2$, whose first terms are

$$-1,\ +4,\ -9,\ +16,\ -25,\ +36, \ldots$$

and are neither all greater nor all smaller than an arbitrary number M when n is large enough.

Definition 1.4. *A sequence $\{a_n\}$ is called* ***irregular*** *(also: oscillating) if it is neither convergent nor divergent. In this case we say that $\lim_{n\to+\infty} a_n$ does not exist.*

One last remark. We have seen that the sequence $b_n = (-1)^n n^2$ is irregular, therefore $\lim b_n$ *does not exist.* However, if we take the absolute values of b_n we find the sequence $\{n^2\}$, which is divergent to $+\infty$. In this case some references say that the limit of the sequence b_n is infinity (without a sign!) and write

$$\lim_{n\to+\infty} (-1)^n n^2 = \infty.$$

That is, in general if $\lim |a_n| = +\infty$ we could say that $\{a_n\}$ *diverges* to ∞ (without a sign). However, in some cases the use of the symbol ∞ may turn out to be rather tricky for an unskilled user. We shall therefore always specify the sign of the infinity we are dealing with, and avoid the use of the symbol ∞ without a sign.

3.1.5 Uniqueness of the limit

The asymptotic behaviour of a sequence $\{a_n\}$ must follow one of three possibilities.

— It converges ($\lim_{n\to+\infty} a_n = A$).

— It diverges ($\lim_{n\to+\infty} a_n = -\infty$ or $+\infty$).

— It is irregular ($\lim_{n\to+\infty} a_n$ does not exist).

Let us suppose we are in one of the first two cases, that is one of the cases in which a limit exists. Is it possible that a sequence tends to two different limits? For example, is it possible to find a sequence $\{a_n\}$ such that

$$\lim_{n\to+\infty} a_n = 1 \quad \text{and} \quad \lim_{n\to+\infty} a_n = 6?$$

The fact that $\{a_n\}$ converges to 1 implies that $a_n < 2$, at least when n is large enough; while the fact that $\{a_n\}$ converges to 6 implies that $a_n > 5$, at least when n is large enough. The two properties contrast with each other, which means they cannot both be true at the same time. The same argument can easily be generalized, and we can conclude that:

Theorem 1.1. *If a sequence $\{a_n\}$ has a limit, it is unique.*

3.1.6 Limits of elementary sequences

We invite the reader to check the following results. This can be done with the help of definitions 1.2 and 1.3.

- *Limit of a power sequence.* For the sequence

$$a_n = n^\alpha$$

we have

$$n^\alpha \to \begin{cases} +\infty & \text{if } \alpha > 0 \\ 1 & \text{if } \alpha = 0 \\ 0^+ & \text{if } \alpha < 0. \end{cases}$$

- *Limit of a geometric sequence.* For the sequence

$$a_n = q^n$$

we have

$$q^n \to \begin{cases} +\infty & \text{if } q > 1 \\ 1 & \text{if } q = 1 \\ 0 & \text{if } |q| < 1 \end{cases}$$

For $q \leq -1$ the sequence is irregular.

Let us prove, for example, that q^n is infinitesimal for $|q| < 1$. This corresponds to showing that, for any $\varepsilon > 0$, when n is large enough we have

$$|q|^n < \varepsilon.$$

If we now consider logarithms with base $|q|$ we get, since $|q| < 1$, that the previous condition holds whenever

$$n > \ln_{|q|} \varepsilon.$$

- *Limit of a logarithmic sequence.* Let us consider the sequence

$$a_n = \log_a n$$

in which the base a must be positive and different from 1. We have

$$\log_a n \to \begin{cases} +\infty & \text{if } a > 1 \\ -\infty & \text{if } a < 1. \end{cases}$$

3.2 Limits of functions

We have defined the limit operation for sequences. This makes it quite easy to extend the operation to all functions.

However, there is a difference between the two cases. A sequence depends on a variable n which only takes integer values, and the limit operation describes its asymptotic behaviour when n *tends to* $+\infty$. That is, for a sequence it does not make sense to talk about a *limit as* n *tends to* 12*, or* 1223 *or* 1.000.000. On the contrary,

the limit operation for a function f, which is typically defined on an interval (a, b), is useful for describing the behaviour of its graph in a neighbourhood of a, or b, or any other point c between a and b. As a consequence, the possible cases to be considered will be much more varied than before. Near to the point a, we shall only be able to move in a right neighbourhood of a; we shall talk about the *right-hand limit* of f at a, that is the limit of f as x tends to a from the right. Analogously, we shall talk about the *left-hand limit* of f at b; and we shall talk about *both* the right-hand limit and the left-hand limit of f at c. Finally, a could also coincide with $-\infty$, or b with $+\infty$, and in those cases we shall talk about limits as x tends to $-\infty$ or to $+\infty$.

3.2.1 Right-hand limit

Let us examine the case of a function which is defined in a *right neighbourhood* of a real number c, that is an interval $(c, c + h)$, with $h > 0$. We want to emphasize a fact which may appear a minor detail but is not: *it is not relevant whether the function is defined at c or not.* And if the function were defined at c, its value at c would be completely irrelevant as far as the definition of the limit is concerned.

Let f be a function, defined in the interval $(c, c + h)$.

Definition 2.1. *We say that the* ***right-hand limit*** *of f as x tends to c (that is: the limit of f at c, as x tends to c from the right) is L (finite or infinite) if, for any sequence $\{x_n\}$ of points belonging to the domain of f, such that $x_n > c$ for all values of n, which is convergent to c, the sequence $\{f(x_n)\}$ of the images of those points tends to L.*

In this case we write that

$$\lim_{x \to c^+} f(x) = L$$

or

$$f(x) \to L \quad \text{as} \quad x \to c^+.$$

More precisely: if L is a real number every sequence $\{f(x_n)\}$ *converges to* L, while if $L = +\infty$ or $-\infty$ every sequence $\{f(x_n)\}$ *diverges to* $+\infty$ or $-\infty$ respectively.

3.2.2 Left-hand limit. Limit

If f is a function which is defined in a left neighbourhood $(c - h, c)$ of c, we will choose sequences $\{x_n\}$ which are convergent to c, but whose terms are all smaller than c. In this case we write

$$\lim_{x \to c^-} f(x) = L$$

or

$$f(x) \to L \quad \text{as} \quad x \to c^-,$$

where L can still be finite or infinite.

If f is defined both on the right and on the left of a real number c, i.e. it is defined in an open interval $(c - h, c + h)$ containing c, with the possible exclusion of the point c itself, we can give the definition of a bilateral limit.

Definition 2.2. *If*

$$\lim_{x\to c^-} f(x) = \lim_{x\to c^+} f(x) = L$$

we say that the ***bilateral limit*** *(from now on, simply*: *the* ***limit****) of* $f(x)$ *as* x *tends to* c *is* L *(finite or infinite) and write*

$$\lim_{x\to c} f(x) = L \quad \text{or} \quad f(x) \to L \text{ as } x \to c.$$

• *Vertical asymptotes.* When $f(x) \to +\infty$ or $f(x) \to -\infty$ as x tends to c from the right and/or from the left, we say that the straight line having equation $x = c$ is a *vertical asymptote* for the graph of f. For example, given the function $g(x) = 1/(x-3)$, the straight line with equation $x = 3$ is a vertical asymptote for its graph.

3.2.3 Limit as $x \to +\infty, -\infty$

Let now f be defined in an interval which is unbounded to the right and/or to the left. In the first case, it is possible to define the limit as x tends to $+\infty$; in the second, the limit as x tends to $-\infty$.

There is nothing substantially new, here. If we deal with a limit as x tends to a real number c, we have to consider sequences $\{x_n\}$ which *converge to* c. If we deal with a limit as x tends to $+\infty$ $(-\infty)$, we shall choose sequences which *diverge to* $+\infty$ $(-\infty)$ and leave everything else unchanged in the definition. We shall write

$$\lim_{x\to+\infty} f(x) = L \qquad \left(\lim_{x\to-\infty} f(x) = L\right)$$

and L can be finite or infinite, as usual. For example, the possible asymptotic behaviour of a function with a *finite* limit L when $x \to +\infty$ is shown in the two following graphs.

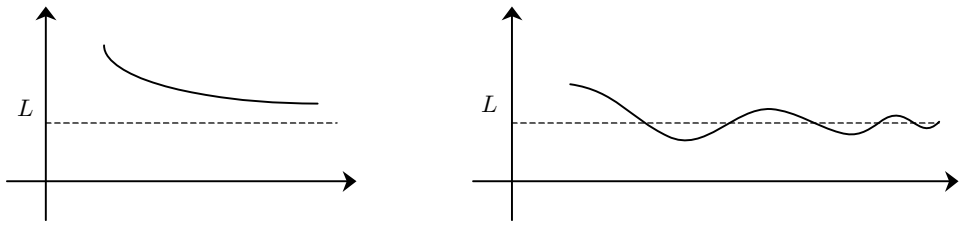


Figure 3.2. Horizontal asymptotes

• *Horizontal asymptotes.* As we can see from the graphs, when $f(x) \to L$ (finite) as $x \to +\infty$ (or $-\infty$) the straight line with equation $y = L$ has a special role, and takes the name of *horizontal asymptote.*

In order to help readers find their way among the various definitions of a limit, we present a collection of graphs which show some typical situations, with the corresponding description in terms of limits. We have:

(a) $\lim\limits_{x\to 2} f(x) = 1$

(b) $\lim\limits_{x\to 2^-} f(x) = 1, \quad \lim\limits_{x\to 2^+} f(x) = 3$

(c) $\lim\limits_{x\to 2^-} f(x) = -\infty, \quad \lim\limits_{x\to 2^+} f(x) = +\infty$

(d) $\lim\limits_{x\to -\infty} f(x) = +\infty$

(e) $\lim\limits_{x\to +\infty} f(x) = 3$

(f) $\lim\limits_{x\to +\infty} f(x)$ does not exist.

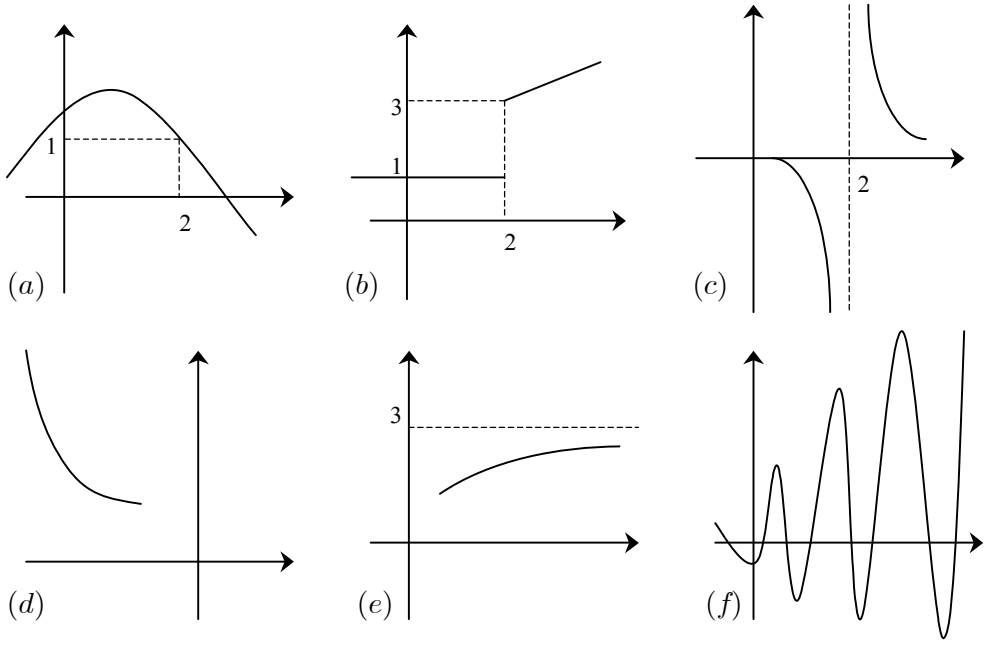


Figure 3.3. Graphs corresponding to different kinds of limits

Definitions 2.3. *A function f which tends to 0 as $x \to c$ (finite or infinite) is called an **infinitesimal** for $x \to c$. A function f which tends to $+\infty$ (or $-\infty$) as $x \to c$ (finite or infinite) is called an **infinity** for $x \to c$.*

Note that it is compulsory to specify the point to which x tends. For example, the function $f(x) = x$ is an infinitesimal when $x \to 0$, while it is an infinity when $x \to +\infty$.

Theorem 2.1 (Uniqueness of the limit). *If a function f has a limit as x tends to c (finite or infinite), this limit is unique.*

This result is a direct consequence of the analogous theorem for sequences, and of the definition of a limit. If a function f had two different limits l, l' when $x \to c$, it

would be possible to find two sequences $\{x_n\}$ and $\{x'_n\}$, both convergent to c, such that the corresponding values $\{f(x_n)\}$ and $\{f(x'_n)\}$ converge to l and l' respectively. But this is not possible, as the definition of limit requires that for all sequences $\{x_n\}$ which are convergent to c, *all* the corresponding sequences $\{f(x_n)\}$ converge to *the same limit.*

3.3 Existence of the limit

3.3.1 Limit of a monotonic sequence

For monotonic sequences the following (important) theorem holds.

Theorem 3.1. *A monotonic sequence is always regular: it either converges or diverges.*

We can state a more precise result: monotonicity and boundedness imply convergence.

A sequence $\{a_n\}$ which is increasing (at least when n is large enough), and bounded from above, converges. The same holds for a sequence which is decreasing (at least when n is large enough), and bounded from below.

The proof of these results depends on the property of the field of real numbers which we called completeness (see Chapter 1).

3.3.2 Limit of a monotonic function

A similar result also holds for monotonic functions.

Theorem 3.2. *Let $a < c < b$, and let f be monotonic on the interval (a, b). Then the two limits*

$$\lim_{x \to c^+} f(x) \quad \textit{and} \quad \lim_{x \to c^-} f(x)$$

exist and are finite, and also the two limits

$$\lim_{x \to a^+} f(x) \quad \text{e} \quad \lim_{x \to b^-} f(x)$$

exist (finite or infinite).

Let us consider the graphs of the two increasing functions in figure 4.

In the first case, the right-hand limit as $x \to c$ is coincident with $f(c)$, the value of the function at the point c, and this coincides with the *minimum* of f in the interval $[c, b)$. In the second graph, the left-hand limit is coincident with $f(c)$, and this turns out to be the *maximum* of f in the interval $(a, c]$. In the first graph the left-hand limit is strictly less than $f(c)$, while in the second graph the right-hand limit is strictly greater than $f(c)$. In any case we have

$$\lim_{x \to c^-} f(x) \leq f(c) \leq \lim_{x \to c^+} f(x).$$

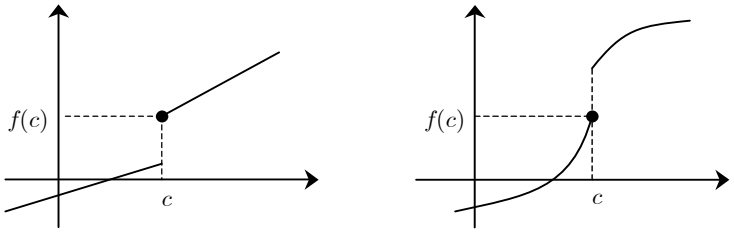


Figure 3.4. Limits of monotonic functions

A similar analysis is possible when the function f is decreasing.

A case in which it is easy to identify the limit occurs when the limit is evaluated at the extremes a or b of the interval, and the function is unbounded in a neighbourhood of those points. The limit (as $x \to a^+$, or $x \to b^-$) is then equal to infinity, with a sign which depends on the kind of monotonicity of f. Graphical examples will explain this better than words: let us examine figure 5.

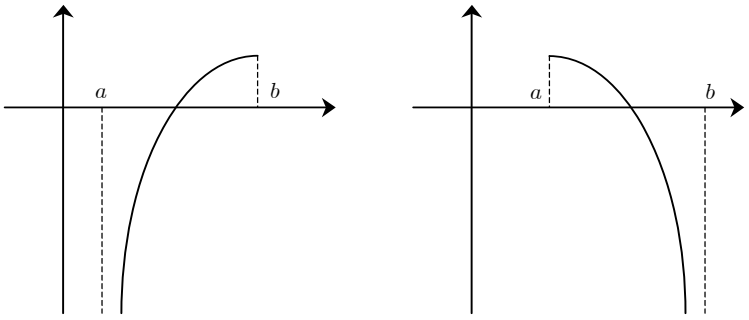


Figure 3.5. Monotonic and unbounded functions

In the first case f is increasing and unbounded in a neighbourhood of a. It follows that

$$\lim_{x \to a^+} f(x) = -\infty$$

and the function has a vertical asymptote with equation $x = a$. In the second case f is decreasing and unbounded in a neighbourhood of b. It follows that

$$\lim_{x \to b^-} f(x) = -\infty$$

and the function has a vertical asymptote with equation $x = b$.

3.3.3 Limits of elementary functions

The existence of limits for monotonic functions allows us to compute the limits of elementary functions quite easily. Indeed, if we work with a monotonic function we

can be sure that the limit we are trying to compute exists, therefore we can calculate it by choosing one *particular* sequence x_n. And obviously we shall choose the most convenient one. For example, when $x \to +\infty$ we can choose $x_n = n$, so that we can use for functions all the results we have already obtained for sequences.

We invite the reader to check the correctness of the following results, with the help of the graphs we saw in Chapter 2.

1. *Powers.* If $x_0 > 0$ we have

$$\lim_{x \to x_0} x^\alpha = x_0^\alpha$$

In some particular cases we also accept $x_0 < 0$; for example when $\alpha = 1, 2, ..., n$, that is, a positive integer. Moreover:

$$\lim_{x \to 0^+} x^\alpha = \begin{cases} 0 & \text{if } \alpha > 0 \\ 1 & \text{if } \alpha = 0 \\ +\infty & \text{if } \alpha < 0 \end{cases}, \qquad \lim_{x \to +\infty} x^\alpha = \begin{cases} +\infty & \text{if } \alpha > 0 \\ 1 & \text{if } \alpha = 0 \\ 0 & \text{if } \alpha < 0. \end{cases}$$

The limit as $x \to 0^+$ can be computed by choosing the sequence $x_n = 1/n$. In the case $\alpha < 0$, we have a horizontal asymptote with equation $y = 0$ and a vertical asymptote with equation $x = 0$. When $\alpha = 1, 2, ..., n$, a positive integer, we can also compute the limit as x tends to $-\infty$:

$$\lim_{x \to -\infty} x^n = \begin{cases} +\infty & \text{for } n \text{ even} \\ -\infty & \text{for } n \text{ odd} \end{cases}$$

2. *Exponentials.* We have

$$\lim_{x \to x_0} a^x = a^{x_0}$$

Moreover,

$$\lim_{x \to -\infty} a^x = \begin{cases} 0 & \text{if } a > 1 \\ 1 & \text{if } a = 1 \\ +\infty & \text{if } a < 1 \end{cases}, \qquad \lim_{x \to +\infty} a^x = \begin{cases} +\infty & \text{if } a > 1 \\ 1 & \text{if } a = 1 \\ 0 & \text{if } a < 1 \end{cases}$$

We have a horizontal asymptote with equation $y = 0$, when $x \to +\infty$ in the case $a < 1$ and when $x \to -\infty$ in the case $a > 1$.

3. *Logarithms.* We have

$$\lim_{x \to x_0} \log_a x = \log_a x_0 \quad (x_0 > 0)$$

Moreover,

$$\lim_{x \to 0^+} \log_a x = \begin{cases} -\infty & \text{if } a > 1 \\ +\infty & \text{if } a < 1 \end{cases}, \qquad \lim_{x \to +\infty} \log_a x = \begin{cases} +\infty & \text{if } a > 1 \\ -\infty & \text{if } a < 1 \end{cases}$$

We have a vertical asymptote with equation $x = 0$.

4. *Trigonometric functions.* For every real number x_0,

$$\lim_{x\to x_0} \sin x = \sin x_0, \quad \lim_{x\to x_0} \cos x = \cos x_0.$$

The limits

$$\lim_{x\to-\infty} \sin x, \ \lim_{x\to+\infty} \sin x \qquad \text{and} \qquad \lim_{x\to-\infty} \cos x, \ \lim_{x\to+\infty} \cos x$$

do not exist: the sine and cosine functions indefinitely oscillate as x tends to infinity, and the amplitude of their oscillations is equal to 2.

As we shall see in Chapter 4, the above examples show that *all elementary functions are continuous, at every point of their domain.*

3.4 The number e

An important sequence, also in economical and financial applications, is

$$e_n(\delta) = \left(1 + \frac{\delta}{n}\right)^n,$$

where δ is a real number.

An interesting particular case takes place when $\delta = 1$. The sequence

$$e_n(1) = \left(1 + \frac{1}{n}\right)^n$$

turns out to be increasing and bounded from above, therefore according to theorem 3.1 it is convergent. Its limit is denoted by the letter "e" and is known as *Napier's number*; it is taken as the base of the so-called natural logarithms (also called *Napier* logarithms, and denoted by ln). More precisely, we define:

$$\boxed{e := \lim_{n\to+\infty} \left(1 + \frac{1}{n}\right)^n}$$

Napier's number is irrational, and its value is approximately

$$e = 2.71828...$$

As the sequence $e_n(1)$ is monotonic increasing, for every $n \geqslant 1$ we have

$$\left(1 + \frac{1}{n}\right)^n < e.$$

• (⇛ **Chapter 11**) Let us try to justify our statements about the convergence of $e_n(\delta)$ using a financial rationale. We consider a unitary investment: 1 Euro. Let δ

be the interest rate (that is, the interest produced by one Euro in one year). After one year, the outcome of our investment (called the *accumulated amount*) is $1+\delta$.

The third paragraph of article 821 of the Italian civil code specifies that interests (jurists use the picturesque name of *civil fruits* for them) mature day by day. However, it seems evident that both contracting parties have good reasons for not paying/receiving interests on a daily basis. In practice, payments take place less frequently: usually, financial contracts provide for special dates called *maturities*, at which the total interest calculated up to that time is paid.

Let us suppose that the payment of interest occurs only once a year, at the end of the year: the accumulated amount is $e_1 = 1+\delta$. If the payment occurs every six months, the accumulated amount after six months is $1+\dfrac{\delta}{2}$ and at the end of one year it is

$$\left(1+\frac{\delta}{2}\right)\left(1+\frac{\delta}{2}\right) = 1+\delta+\frac{\delta^2}{4}$$

This expression can easily be interpreted: the first addendum is the Euro we invested, the second addendum represents the interest produced during the year, the third is given by the interest $\frac{\delta}{2}$, produced in the second half of the year, on the interest $\frac{\delta}{2}$ accumulated during the first half. We can understand that, if accumulation takes place n times a year (for example: when $n=4$, every three months, as banks do with our accounts; when $n=12$, every month, as happens with our credit cards), the accumulated amount at the end of the year is

$$e_n(\delta) = \left(1+\frac{\delta}{n}\right)^n.$$

Let us try to understand what happens when $n \to +\infty$, both because this is a good approximation of what should happen according to article 821 of the Italian civil code, and because the hypothesis of *continuous accumulation of interest* is quite common in practice. In this case the interest rate is called the *instant rate of interest* (or *force of interest*), to underline the fact that interests are instantaneously produced (and accumulated).

We note now that the sequence $\{e_n\}$ is monotonic increasing, at least when $\delta > 0$:

$$e_{n+1}(\delta) > e_n(\delta).$$

This is obvious, according to the scheme we outlined above, because when the number of maturities increases new interest is added more frequently to the previous amounts, and therefore the accumulated amount at the end of one year increases.

Let us finally see why $e_n(\delta)$ is also bounded from above; in particular, we shall check that $e_n(1) < 4$.

Suppose first that $0 < \delta < 1$. We show that it is possible to find an upper bound, or "majorant", for $e_n(\delta)$ using a simple financial argument. Let us call "first order interest" the interest which is produced by the initial principal, "second order interest" that which is produced by the first order interest, "third order interest" that which is produced by the second order interest, and so on. The first order

interest on one Euro amounts to δ, the second order interest amounts to *less than* δ^2, the third order interest to *less than* δ^3, and so on, therefore[2]

$$e_n(\delta) < 1 + \delta + \delta^2 + \cdots + \delta^n = \frac{1-\delta^{n+1}}{1-\delta} < \frac{1}{1-\delta}. \tag{3.5}$$

Suppose now that $1 \leq \delta < 2$. If we consider an even number $k = 2n$ of fractions of a year, we have

$$\left(1 + \frac{\delta}{2n}\right)^{2n} = \left(1 + \frac{\delta/2}{n}\right)^{2n} = \left(1 + \frac{\delta/2}{n}\right)^{n} \left(1 + \frac{\delta/2}{n}\right)^{n}.$$

As $1 \leq \delta < 2$, it follows that $1/2 \leq \delta/2 < 1$; therefore, using formula (3.5) with $\delta/2$ instead of δ, we can deduce that each of the factors on the right-hand side is less than $1/(1-\delta/2)$. This implies that

$$e_{2n}(\delta) = \left(1 + \frac{\delta}{2n}\right)^{2n} < \frac{1}{1-\delta/2} \cdot \frac{1}{1-\delta/2} = \left(\frac{1}{1-\delta/2}\right)^{2}.$$

This can be extended to all values of δ such that $1 \leq \delta < 2$. In particular, when $\delta = 1$ we deduce that

$$e_{2n}(1) < \left(\frac{1}{1-1/2}\right)^{2} = 4.$$

As the sequence $e_n(\delta)$ is monotonic, this limitation also holds for all its terms with an odd index.

A similar argument can be used for all other positive values of δ, and this concludes our financial reasoning.

Actually, it can be proved that

$$e_n(\delta) = \left(1 + \frac{\delta}{n}\right)^{n} \to e^{\delta} \quad \text{as } n \to +\infty, \text{ for all } \delta \in \mathbb{R}$$

In the case of an investment for t years, we have

$$e_n(\delta t) = \left(1 + \frac{\delta t}{n}\right)^{n}$$

and therefore

$$e_n(\delta t) \to e^{\delta t} \quad \text{as } n \to +\infty.$$

Let us consider an example. We invest a unit principal for $t = 2$ years, at the instant rate of interest $\delta = 10\%$. The accumulated amount turns out to be $e^{2 \cdot 10\%} = e^{0.2} \simeq 1.2214$. Using simple interest, the accumulated amount would only be 1.2: the difference is due to the continuous accumulation.

[2] Remember the formula for the sum of the first terms of a geometric progression, which was found in section 4.3 of the first Chapter.

We remark that the expression $e^{\delta t}$ appears in many formulae, such as Black and Scholes' formula, which are currently used all over the world to fix the price of financial options, and is therefore very common and relevant.

We can also show that if $\{x_n\}$ is any divergent sequence (not only to $+\infty$, but also to $-\infty$) and δ is a real number, the formula

$$\lim_{n\to+\infty}\left(1+\frac{\delta}{x_n}\right)^{x_n}=e^{\delta}$$

holds. This implies that for any real number δ we have

$$\boxed{\lim_{x\to+\infty}\left(1+\frac{\delta}{x}\right)^{x}=e^{\delta},\qquad\lim_{x\to-\infty}\left(1+\frac{\delta}{x}\right)^{x}=e^{\delta}.}\tag{3.6}$$

For example:

$$\left(1-\frac{5}{n^2}\right)^{n^2}=\left(1+\frac{-5}{n^2}\right)^{n^2}\to e^{-5}.$$

3.5 Calculation of limits

3.5.1 The set $\mathbb{R}^*$

In order to deal with limits, real numbers are not sufficient. Indeed, we need to introduce the two "mysterious objects" $+\infty$ and $-\infty$.

Definition 5.1. *The set* $\mathbb{R}\cup\{-\infty,+\infty\}$ *is denoted by the symbol* $\mathbb{R}^*$ (which reads "ar star").

Calculations of limits must be performed in the set $\mathbb{R}^*$. Let us therefore try to get acquainted with this new set, starting with a geometric image of it. The straight line is a geometric image of the set $\mathbb{R}$, in the sense that there is a one-to-one correspondence between the points of the straight line and the real numbers, but there is an infinite number of other sets which are in one-to-one correspondence with the set $\mathbb{R}$. And we can also find a set of points which is in one-to-one correspondence with our set $\mathbb{R}\cup\{-\infty,+\infty\}$. Let us take an oriented straight line, consider the point O which corresponds to 0 and construct a half-circumference which is tangent to the straight line at O. We then call AB the diameter which is parallel to the straight line, and call C the centre of the half-circumference. Let us consider a half-line, with origin at C, which intersects both the straight line and the half-circumference. We can see that all the points P on the straight line are in one-to-one correspondence with the points P' on the half-circumference, except for A and B. It is now easy to get a geometric image of $\mathbb{R}^*$: it is the *half-circumference, including the two points* A *and* B. The symbol $-\infty$ corresponds to the point A, and $+\infty$ corresponds to B.

We note that $\mathbb{R}^*$ is an *ordered set*: it is enough to impose that for every real number x we have $-\infty<x<+\infty$. We can also introduce a *partial arithmetic structure*

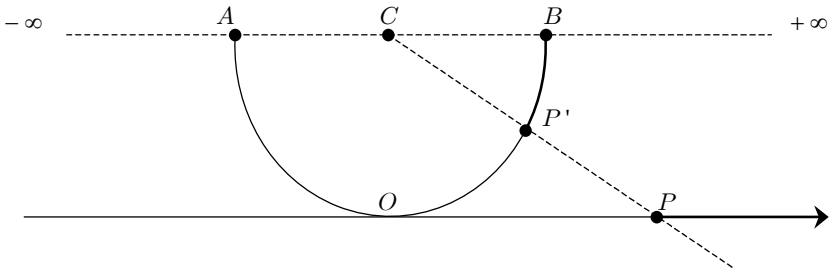


Figure 3.6. Geometric image of $\mathbb{R}^*$

into $\mathbb{R}^*$: indeed we can extend the operations of sum and product to the symbols $+\infty, -\infty$, as we can see from the following table (A denotes any real number).

$(+\infty)+(+\infty)=+\infty$	$(-\infty)+(-\infty)=-\infty$	$A+(\pm\infty)=\pm\infty$
$(\pm\infty)\cdot(\pm\infty)=+\infty$	$(\pm\infty)\cdot(\mp\infty)=-\infty$	$A\neq 0, \quad A\cdot\infty=\infty$
$\dfrac{A}{\infty}=0$	$\dfrac{\infty}{A}=\infty$	$A\neq 0, \quad \dfrac{A}{0}=\infty$

In the cases for which the sign of the infinity is not specified, its sign can usually be determined by looking at the sign of A and considering the good old "rule of signs": $+\cdot+=+$, $+\cdot-=-$, $-\cdot-=+$. And using the definitions which appear in the table, we can check that the main properties of operations still hold.

The arithmetic structure of $\mathbb{R}^*$ is only partial, because some operations cannot be defined. They are the operations we can denote by the symbols

$$(+\infty)+(-\infty), \quad 0\cdot(\pm\infty), \quad 0/0, \quad (\pm\infty)/(\pm\infty). \tag{3.7}$$

3.5.2 Limits and algebraic operations

Consider two functions f, g defined on the same set[3]. We can construct the sum, the product and (if $g\neq 0$) the quotient function

$$f+g \qquad f\cdot g \qquad f/g.$$

If we know the limit of f and g as $x\to c$ (with c finite or infinite), can we compute the limit of $f+g$, $f\cdot g$, f/g ?

The answer is nearly always in the affirmative. In particular, if $L, M\in\mathbb{R}^*$, remembering the partial arithmetic structure we have just seen, we can state that if

$$\lim_{x\to c} f(x)=L, \quad \lim_{x\to c} g(x)=M$$

[3] An interval (bounded or unbounded), or a union of intervals.

then

$$\boxed{\lim_{x\to c}[f(x)+g(x)] = L+M, \quad \lim_{x\to c}[f(x)g(x)] = LM, \quad \lim_{x\to c}\frac{f(x)}{g(x)} = \frac{L}{M},}$$

except for the following cases:

for the sum	$L=+\infty,\ M=-\infty$ or vice versa;
for the product	$L=0,\ M=\pm\infty$ or vice versa;
for the quotient	$L=M=0$ or $L=M=\pm\infty$.

These anomalies correspond exactly to the *indeterminate forms* (3.7). This does not mean that when this situation occurs the limit cannot be decided! It just means that anything can happen, as we can see from the following examples. They involve sequences; obviously, the same results and considerations seen with functions apply.

Examples

In the first three examples, the sequence a_n diverges to $+\infty$ while the sequence b_n diverges to $-\infty$. In examples 4, 5, 6, the first sequence is an infinity and the second one is an infinitesimal. In examples 7 and 8 both sequences a_n and b_n diverge.

5.1. $a_n = n^2+1$, $b_n = -n^2$. We have $s_n = a_n + b_n = 1 \to 1$ (it is a constant sequence, with general term equal to 1).

5.2. $a_n = n^2+n$, $b_n = -n^2$. We have $s_n = a_n + b_n = n \to +\infty$.

5.3. $a_n = n^2+(-1)^n$, $b_n = -n^2$. We have $s_n = a_n+b_n = (-1)^n$ which is irregular.

5.4. $a_n = n$, $b_n = 1/n$. We have $p_n = a_n \cdot b_n = 1 \to 1$.

5.5. $a_n = n^2$, $b_n = 1/n$. We have $p_n = a_n \cdot b_n = n \to +\infty$.

5.6. $a_n = n$, $b_n = 1/n^2$. We have $p_n = a_n \cdot b_n = 1/n \to 0$.

5.7. $a_n = 4n+3$, $b_n = n$. We have $q_n = a_n/b_n = 4 + \dfrac{3}{n} \to 4$.

5.8. $a_n = n^2$, $b_n = -n$. We have $q_n = a_n/b_n = -n \to -\infty$.

3.5.3 Limits and inequalities

Let f, g be defined in the same interval I. If $f(x) \le g(x)$ for every $x \in I$ and we apply a limit operation on f and g, this operation does not "destroy" the inequality. To be precise, we can prove the following theorems.

Theorem 5.1 (Comparison). *If $f(x) \le g(x)$ in a neighbourhood of c, and the limits of f and g as $x \to c$ exist, then*

$$\lim_{x\to c} f(x) \le \lim_{x\to c} g(x). \tag{3.8}$$

In the particular case when $f(x) = 0$, we can deduce that if $g(x) \geq 0$ in a neighbourhood of c, and $\lim_{x\to c} g(x)$ exists, then

$$\lim_{x\to c} g(x) \geq 0,$$

which is a property known as the *permanence of sign*. Note that if $g(x)$ is strictly positive, the limit can well be equal to zero; this means that the property does not hold for strong inequalities. For example, $g(x) = 1/x \to 0$ as $x \to +\infty$.

For sequences, theorem 5.1 becomes: *If $a_n \leqslant b_n$, at least when n is large enough, and $\lim_{n\to+\infty} a_n$, $\lim_{n\to+\infty} b_n$ exist, then*

$$\lim_{n\to+\infty} a_n \leqslant \lim_{n\to+\infty} b_n.$$

In particular, if a_n diverges to $+\infty$ also b_n diverges to $+\infty$; and if b_n diverges to $-\infty$ also a_n diverges to $-\infty$.

Proof. If a_n diverges to $+\infty$, for every real number M we have

$$a_n > M$$

when n is large enough. And as $a_n \leqslant b_n$ when n is large enough, we also have that

$$b_n > M$$

when n is large enough, therefore b_n also diverges to $+\infty$. In the other cases the proof is similar. □

Example 5.9. The sequence $s_n = n(2 + \sin n)$ diverges to $+\infty$. In fact

$$\sin n \geq -1 \quad \Longrightarrow \quad 2 + \sin n \geq 1 \quad \Longrightarrow \quad n(2 + \sin n) \geq n \to +\infty.$$

In theorem 5.1, in the case where one of the two functions has a finite limit we can say nothing about the other one. But in the particular case in which a double comparison exists, and two functions "trap" a third one, the theorem allows us to calculate the limit[4]. Precisely:

Theorem 5.4. *If $f(x) \leq g(x) \leq h(x)$ in a neighbourhood of c, and*

$$\lim_{x\to c} f(x) = \lim_{x\to c} h(x) = L,$$

then also

$$\lim_{x\to c} g(x) = L.$$

Example 5.10. We prove that

$$\boxed{\lim_{x\to+\infty} \frac{\sin x}{x} = 0}$$

[4] This form of the comparison theorem is known by many picturesque names. One of them is: the theorem *of the two policemen.*

Indeed $-1 \leq \sin x \leq 1$; therefore when $x > 0$

$$-\frac{1}{x} \leq \frac{\sin x}{x} \leq \frac{1}{x}.$$

The function $\dfrac{\sin x}{x}$ is trapped between $f(x) = -\dfrac{1}{x}$ and $h(x) = \dfrac{1}{x}$, which tend to 0 as $x \to +\infty$.

Example 5.11. We prove that[5]

$$\boxed{\lim_{x \to 0} \frac{\sin x}{x} = 1} \tag{3.9}$$

As $\sin x/x$ is an even function (that is, its graph is symmetric with respect to the y-axis), it is sufficient to compute the limit as $x \to 0^+$. Let us consider the unit circle centred at the origin, the angle x (measured in radians) and the points P, H, A, T as in Figure 7. We can see that

$$\text{area of triangle } OAP \ < \ \text{area of sector } OAP \ < \ \text{area of triangle } OAT. \tag{3.10}$$

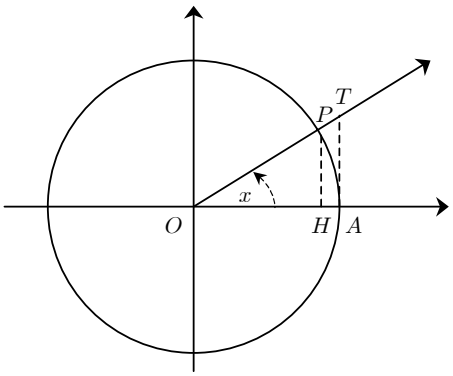


Figure 3.7. $\sin x < x < \tan x$

According to the definition of *radian* the length of the arc AP is x, while according to the definitions of *sine* and *tangent* of an angle we have

$$\overline{HP} = \sin x \quad \text{and} \quad \overline{AT} = \tan x = \frac{\sin x}{\cos x}.$$

Therefore the relation (3.10) becomes

$$\frac{1}{2} \sin x < \frac{1}{2} x < \frac{1}{2} \frac{\sin x}{\cos x}$$

[5] The limit is in the indeterminate form $0/0$.

If we multiply by $2/\sin x > 0$, we get

$$1 < \frac{x}{\sin x} < \frac{1}{\cos x}$$

and therefore

$$\cos x < \frac{\sin x}{x} < 1.$$

The function $\dfrac{\sin x}{x}$ is now trapped between $f(x) = \cos x$ and $h(x) = 1$, and they both tend to 1 as $x \to 0^+$. We can conclude that (3.9) holds.

3.5.4 Change of variable

In the calculation of limits, a change of variable often turns out to be useful. Let us suppose that we want to compute

$$\lim_{x \to c} f[g(x)]$$

and we know that $\lim_{x \to c} g(x) = k$. We would like to set $y = g(x)$ and compute $\lim_{y \to k} f(y)$. When is this possible? That is, when is the formula

$$\lim_{x \to c} f[g(x)] = \lim_{y \to k} f(y) \tag{3.11}$$

true?

The answer is quite simple: *it is certainly true if* $\lim_{y \to k} f(y) = f(k)$,[6] *and it is also true in the case where* k *is equal to* $+\infty$ *or to* $-\infty$.

Examples

5.12. As $\lim_{x \to 2^-} \dfrac{1}{x-2} = -\infty$, we have

$$\lim_{x \to 2^-} e^{1/(x-2)} = \lim_{y \to -\infty} e^y = 0^+.$$

5.13. Let us compute $\lim_{x \to 0} (1 + \delta x)^{1/x}$. With the change of variable $y = 1/x$, we get $y \to +\infty$ when $x \to 0^+$ and $y \to -\infty$ when $x \to 0^-$. Therefore, according to the formulae (3.6):

$$\lim_{x \to 0} (1 + \delta x)^{1/x} = \lim_{y \to \pm\infty} \left(1 + \frac{\delta}{y}\right)^y = e^\delta.$$

5.14. Let us show that

$$\boxed{\lim_{x \to 0} \frac{\ln(1+x)}{x} = 1}$$

[6] I.e., if the function f is *continuous* at k, as we shall learn to say in Chapter 4.

Using the properties of logarithms and the previous example with $\delta = 1$, we have

$$\lim_{x\to 0}\frac{1}{x}\ln(1+x) = \lim_{x\to 0}\ln(1+x)^{1/x} = \ln e = 1.$$

5.15. Let us show that

$$\boxed{\lim_{x\to 0}\frac{e^x-1}{x} = 1}$$

We set $y = e^x - 1$. This gives us $x = \ln(1+y)$, and $x \to 0$ exactly when $y \to 0$. Therefore, using the previous example, we obtain

$$\lim_{x\to 0}\frac{e^x-1}{x} = \lim_{y\to 0}\frac{y}{\ln(1+y)} = 1.$$

3.6 Comparisons

In general, if we want to solve an indeterminate form, an accurate analysis of the "speed" with which the sequences or the functions involved tend to zero or to infinity is necessary. The comparison theorems will turn out to be useful, in order to establish some "hierarchies" among infinities and infinitesimals.

3.6.1 The symbols "o" and "$\sim$"

First of all, let us introduce the two symbols "o" and "$\sim$". Suppose that f, g are two functions which are defined in a neighbourhood of c, and that $g \neq 0$ in that neighbourhood (with the possible exclusion of the point c itself).

Definition 6.1. *If*

$$\lim_{x\to c}\frac{f(x)}{g(x)} = 0$$

*we say that $f(x)$ is $o(g(x))$ as $x \to c$, which reads "$f(x)$ is **little-o** of $g(x)$ as x tends to c".*

The symbol "o" was introduced by the German mathematician Edmund Landau (1877-1938). A very common slight abuse of notation consists in writing:

"$f(x) = o(g(x))$ *as* $x \to c$" instead of "$f(x)$ *is* $o(g(x))$ *as* $x \to c$".

A good intuitive interpretation of the sentence "$f(x)$ is $o(g(x))$ as $x \to c$" is that "$f(x)$ is negligible with respect to $g(x)$ as $x \to c$".

Definition 6.2. *If*

$$\lim_{x\to c}\frac{f(x)}{g(x)} = 1$$

*we say that $f(x) \sim g(x)$ as $x \to c$, which reads "$f(x)$ is **asymptotic** to $g(x)$ as x tends to c".*

A good intuitive interpretation of the sentence "$f(x) \sim g(x)$ as $x \to c$" is that "f and g have the same behaviour as $x \to c$". And if they are two infinities (or two infinitesimals), they tend to infinity (or to zero) with the same speed.

The relation $\sim$ of *asymptotic* expresses an *equivalence* of behaviour with respect to the limit operation. Indeed it satisfies the three properties:

— reflexive: $f(x) \sim f(x)$ as $x \to c$;

— symmetric: $f(x) \sim g(x)$ as $x \to c$ if and only if $g(x) \sim f(x)$ as $x \to c$. Therefore we can say that "two functions f, g are asymptotic" (without specifying which is the numerator and which is the denominator of the fraction);

— transitive: if $f(x) \sim g(x)$ as $x \to c$ and $g(x) \sim h(x)$ as $x \to c$, then $f(x) \sim h(x)$ as $x \to c$.

The relation o of *little-o* satisfies the transitive property, like the relation $<$:

— if $f(x) = o(g(x))$ as $x \to c$ and $g(x) = o(h(x))$ as $x \to c$, then $f(x) = o(h(x))$ as $x \to c$.

The asymptotic symbol is particularly useful for the calculation of limits. As the $\sim$ relation expresses an equivalence of behaviour, in the calculation of the limit of a product or of a quotient we can substitute every function (or sequence; the definitions and properties are the same) with a function which is asymptotic to it and simpler, so that the calculations become easier. Indeed, if $f_1(x) \sim f_2(x)$ as $x \to c$ and $g_1(x) \sim g_2(x)$ as $x \to c$, we can easily see that

$$\lim_{x\to c} f_1(x)\, g_1(x) = \lim_{x\to c} f_2(x)\, g_2(x), \qquad \lim_{x\to c} \frac{f_1(x)}{g_1(x)} = \lim_{x\to c} \frac{f_2(x)}{g_2(x)}.$$

A typical way of showing that $f(x) \sim g(x)$ as $x \to c$ consists of writing $f(x) = g(x)\, h(x)$, with $h(x) \to 1$ as $x \to c$.

Examples

6.1. If $\alpha < \beta$, we have

$$\lim_{x\to+\infty} \frac{x^\alpha}{x^\beta} = \lim_{x\to+\infty} x^{\alpha-\beta} = 0, \qquad \lim_{x\to 0^+} \frac{x^\beta}{x^\alpha} = \lim_{x\to 0^+} x^{\beta-\alpha} = 0.$$

Therefore when $\alpha < \beta$ we can write

$$x^\alpha = o\left(x^\beta\right) \quad \text{as } x \to +\infty, \quad x^\beta = o\left(x^\alpha\right) \quad \text{as } x \to 0^+.$$

This means that among all the powers of x, when $x \to +\infty$ those with a lower exponent are negligible; and when $x \to 0$ those with a higher exponent are negligible.

6.2. When $n \to +\infty$, every polynomial in the variable n is asymptotic to its term of maximum degree: if $a_n = -n^3 + 2n^2 - n + 2005$ we can write $a_n \sim -n^3$; indeed

$$\frac{-n^3 + 2n^2 - n + 2005}{-n^3} = 1 - \frac{2}{n} + \frac{1}{n^2} - \frac{2005}{n^3} \to 1 \tag{3.12}$$

as all the fractions on the right-hand side of formula (3.12) tend to zero.

6.3. When $n \to +\infty$, the quotient of two polynomials in the variable n is asymptotic to the quotient of the terms of maximum degree: for example, we can write

$$\frac{-n^3 + 2n^2 - n + 2005}{3n^4 + n^3 - \pi} \sim \frac{-n^3}{3n^4} = -\frac{1}{3n} \to 0.$$

6.4. As $x \to 0$, we have

$$f(x) = x^3 + 2x^4 - 5x^7 = x^3\left(1 + 2x - 5x^4\right) \sim x^3.$$

This means that in a neighbourhood of $x = 0$ the graph of f is "well-approximated" by the graph of $y = x^3$.

Remark. The two symbols $\sim$ and o are linked by the relation

$$f \sim g \quad (\text{as } x \to c) \quad \textit{if and only if} \quad f = g + o(g) \quad (\text{as } x \to c).$$

Note that in the calculation of the limit of a sum, only the terms which are not negligible with respect to the others are to be considered.

Examples

6.5. As $x \to 0$, from the examples 5.13 and 5.14 we get

$$\begin{aligned} \ln(1+x) &\sim x \\ \ln(1+x) &= x + o(x) \end{aligned} \tag{3.13}$$

$$\begin{aligned} e^x - 1 &\sim x \\ e^x - 1 &= x + o(x) \end{aligned} \tag{3.14}$$

6.6. From the example 5.11, we get

$$\begin{aligned} \sin x &\sim x \qquad\quad \text{as } x \to 0 \\ \sin x &= x + o(x) \quad \text{as } x \to 0. \end{aligned}$$

6.7. As $x \to 0$, we have

$$\cos x = 1 + o(x),$$

that is

$$\lim_{x \to 0} \frac{\cos x - 1}{x} = 0.$$

We obtain this result by multiplying the numerator and the denominator by $\cos x + 1$ (which is $\neq 0$ when x is in a neighbourhood of 0); we can then use the fact that $1 - (\cos x)^2 = (\sin x)^2$ and $\sin x \sim x$ as $x \to 0$ (example 6.6):

$$\lim_{x \to 0} \frac{\cos x - 1}{x} \frac{\cos x + 1}{\cos x + 1} = \lim_{x \to 0} \frac{-(\sin x)^2}{x(\cos x + 1)} = \lim_{x \to 0} \frac{-x}{\cos x + 1} = 0.$$

Note that the two symbols "$\sim$" and "o" can be used with *all functions* and *all sequences*, not only infinities and infinitesimals. We can write $a_n \sim 3$, for example, to say that the sequence a_n converges to 3. And if we want to say that a function f is infinitesimal when $x \to c$, we can write $f(x) = o(1)$.

• *Oblique asymptotes.* If the limit of a function f is $+\infty$ $(-\infty)$ when $x \to +\infty$ $(-\infty)$, the graph of f can have many different aspects. An interesting case takes place when a straight line exists (not parallel to either of the two axes), having equation $y = mx + q$ $(m \neq 0)$, such that when $x \to +\infty$ $(-\infty)$ we have

$$f(x) = mx + q + o(1). \tag{3.15}$$

In this case, we say that the straight line $y = mx + q$ is an *oblique asymptote* for f as $x \to +\infty$ $(-\infty)$.

For example, the function $f(x) = x + 1/x$ has the straight line $y = x$ as oblique asymptote, both when $x \to +\infty$ and when $x \to -\infty$.

In fact, $1/x \to 0$ and $f(x) = x + o(1)$ as $x \to \pm\infty$.

The condition expressed by (3.15) is equivalent to the following pair of conditions, which are often simpler to use:

$$\lim_{x\to+\infty} \frac{f(x)}{x} = m, \qquad \lim_{x\to+\infty} [f(x) - mx] = q \qquad (m \neq 0)$$

when $x \to +\infty$ (and the analogous ones when $x \to -\infty$).

Example 6.8. Let $f(x) = \dfrac{2x^2+3}{x-1}$. When $x \to \pm\infty$,

$$\frac{f(x)}{x} = \frac{2x^2+3}{x^2-x} \to 2$$

This means that if there is an oblique asymptote, its slope is 2. And as

$$f(x) - 2x = \frac{2x^2+3}{x-1} - 2x = \frac{3+2x}{x-1} \to 2$$

the oblique asymptote does exist, when $x \to \pm\infty$, and its equation is $y = 2x + 2$.

3.6.2 The hierarchy of infinities

Let us go now into more detail. Suppose that f, g are two functions which are defined in a neighbourhood of c (with the possible exclusion of the point c itself), and that f, g are infinities when $x \to c$. If we consider the limit of the quotient f/g, we have four possible cases:

$$\lim_{x\to c} \frac{f(x)}{g(x)} = \begin{cases} 0 & (a) \\ L \text{ finite and not null} & (b) \\ +\infty \text{ or } -\infty & (c) \\ \text{does not exist} & (d) \end{cases}$$

We introduce the following terminology.

$(a) \implies f$ is an *infinity of lower order* than g, as $x \to c$. In this case, obviously $f(x) = o(g(x))$ when $x \to c$.

$(b) \implies f$ is an *infinity of the same order* of g, as $x \to c$. In particular, if $L = 1$ obviously $f(x) \sim g(x)$ when $x \to c$. In general, we have $f(x) \sim Lg(x)$ when $x \to c$.

$(c) \Longrightarrow f$ is an *infinity of higher order* than g, as $x \to c$. In this case, obviously $g(x) = o(f(x))$ when $x \to c$.

$(d) \Longrightarrow f$ and g are *not comparable.* [7]

Example 6.1 shows that all powers x^α, when $x \to +\infty$, can be "ordered" with respect to their exponent. Analogously, exponentials with different bases can be ordered with respect to their base. For example, let us check that

$$2^x = o(3^x) \quad \text{as } x \to +\infty$$

Indeed we have

$$\frac{2^x}{3^x} = \left(\frac{2}{3}\right)^x \to 0.$$

We can also compare the speeds of the three families of functions: exponentials, powers and logarithms. The following important result holds.

Theorem 6.1. *Every exponential infinity is of higher order than every power infinity*; *every power infinity is of higher order than every logarithmic infinity. That is, for every* $\alpha > 1, \beta > 0, \gamma > 0$

$$\boxed{\lim_{x\to+\infty} \frac{\alpha^x}{x^\beta} = +\infty, \quad \lim_{x\to+\infty} \frac{x^\beta}{(\ln x)^\gamma} = +\infty.}$$

We prove the first result for sequences, using the criterion which is stated in the following theorem. The second result can be deduced from the first one, using the change of variable $y = \ln x$.

Theorem 6.2 (Criterion of the ratio for sequences). *Let* $\{r_n\}$ *be a sequence with positive general term, and suppose that*

$$\lim_{n\to+\infty} \frac{r_{n+1}}{r_n} = L,$$

where L *can also be equal to* $+\infty$. *If*

$$\begin{cases} L > 1 & \text{then } r_n \to +\infty \\ 0 \le L < 1 & \text{then } r_n \to 0 \end{cases}$$

If $L = 1$, *nothing can be said.*

We apply the criterion to the sequence $r_n = \dfrac{\alpha^n}{n^\beta}$.

We have

$$\frac{r_{n+1}}{r_n} = \frac{\alpha^{n+1}/(n+1)^\beta}{\alpha^n/n^\beta} = \alpha\left(\frac{n}{n+1}\right)^\beta \to \alpha > 1.$$

therefore $r_n \to +\infty$.

---

[7] It may happen that $\lim_{x\to c} \frac{f(x)}{g(x)}$ does not exist, but $\lim_{x\to c} \left|\frac{f(x)}{g(x)}\right| = +\infty$. In this case we can say that $\lim_{x\to c} \frac{f(x)}{g(x)} = \infty$ (without sign), and we consider f to be an infinity of higher order than g.

• *The sequence* $\{n!\}$. There exist sequences which are infinities of higher order than any exponential sequence $\{\alpha^n\}$ (with $\alpha > 1$). One example is given by the sequence $\{n!\}$. Indeed:

$$\lim_{n\to+\infty} \frac{\alpha^n}{n!} = 0. \tag{3.16}$$

The result (3.16) can be checked by using the criterion of the ratio: if we set $r_n = \dfrac{\alpha^n}{n!}$, we have

$$\lim_{n\to+\infty} \frac{r_{n+1}}{r_n} = \lim_{n\to+\infty} \frac{\alpha^{n+1}}{(n+1)!}\frac{n!}{\alpha^n} = \lim_{n\to+\infty} \frac{\alpha}{n+1} = 0 \ (<1).$$

• *Malthus' law*[8]. One of the reasons for birth control is the famous (and probably false) Malthusian law, which states that the resources for consumption r_n grow linearly with time n:

$$r_n = a + bn, \qquad \text{with } a, b > 0$$

while the population p_n grows exponentially:

$$p_n = AB^n, \qquad \text{with } A > 0 \text{ and } B > 1.$$

The amount of resources for each individual after n years comes out to be $\dfrac{bn+a}{AB^n}$, and when $n \to +\infty$ we have

$$\frac{bn+a}{AB^n} \sim \frac{b}{A} \cdot \frac{n}{B^n} \to 0$$

We can deduce Malthus' argument: we must control births, because otherwise the amount of resources for each individual will tend to vanish.

3.6.3 The hierarchy of infinitesimals

Suppose now that f, g are two functions which are defined in a neighbourhood of c, and that $g \neq 0$ in that neighbourhood (with the possible exclusion of the point c itself). Suppose that f, g are infinitesimals when $x \to c$. To make things simpler, let us suppose that f, g have a definite sign (either positive or negative) in a neighbourhood of c.

If we consider the limit of the quotient f/g, we have four possible cases:

$$\lim_{x\to c} \frac{f(x)}{g(x)} = \begin{cases} 0 & (a) \\ L \text{ finite and not null} & (b) \\ +\infty \text{ or } -\infty & (c) \\ \text{does not exist} & (d) \end{cases}$$

We introduce the following terminology.

[8] Thomas Robert Malthus (1766-1834), English economist.

$(a) \implies f$ is an *infinitesimal of higher order* than g, as $x \to c$. In this case $f(x) = o(g(x))$ when $x \to c$.

$(b) \implies f$ is an *infinitesimal of the same order* of g, as $x \to c$. In particular, if $L = 1$ obviously $f(x) \sim g(x)$ when $x \to c$. In general, $f(x) \sim Lg(x)$ when $x \to c$.

$(c) \implies f$ is an *infinitesimal of lower order* than g, as $x \to c$. In this case $g(x) = o(f(x))$ when $x \to c$.

$(d) \implies f$ and g are *not comparable.*[9]

Note that if f is an infinity of higher (lower) order than g, then $1/f$ and $1/g$ are infinitesimals with a definite sign and $1/f$ is an infinitesimal of higher (lower) order than $1/g$. Therefore the names we chose are perfectly consistent.

Let us check this for sequences. Let $\{a_n\}, \{b_n\}$ be two infinities, and let the first one be an infinity of lower order than the second: $a_n = o(b_n)$. If we consider their reciprocals $\{1/a_n\}, \{1/b_n\}$, which are two infinitesimals, the first one is an infinitesimal of lower order than the second. It is sufficient to note that

$$\frac{1/a_n}{1/b_n} = \frac{b_n}{a_n} = \frac{1}{a_n/b_n}$$

and this diverges, because the denominator tends to zero (with a definite sign).

For example, from $n^3 = o(2^n)$ we can deduce that the infinitesimal $\{2^{-n}\}$ is of a higher order than $\{n^{-3}\}$; from $\ln n = o(\sqrt{n})$ we can deduce that the infinitesimal $\{1/\ln n\}$ is of a lower order than $\{1/\sqrt{n}\}$.

3.7 Exercises

3.1. Which of the following sequences are infinities? And which are infinitesimals?

$$a_n = -\frac{1}{n^2}, \qquad b_n = \sqrt[3]{-n}, \qquad c_n = 4^{-\sqrt{n}}.$$

3.2. Compute the limits:

$$(a) \lim_{x\to-\infty} \ln\frac{x+1}{x+2}, \quad (b) \lim_{x\to+\infty} e^{1/(x-1)}, \quad (c) \lim_{x\to1^+} e^{1/(x-1)}.$$

3.3. Place the following sequences in increasing order of infinity:

$$2^n, \quad n^3, \qquad n!, \qquad (\ln n)^{2005}, \qquad \sqrt[50]{n}, \qquad n^2 \ln n.$$

3.4. Determine the behaviour of the following sequences:

$$a_n = \frac{2\ln n - 3\sqrt{n}}{\sqrt[3]{n}+5}, \qquad b_n = \frac{n^2+3}{2^n+3n}.$$

[9] If $\lim \frac{f(x)}{g(x)}$ does not exist but $\lim \left|\frac{f(x)}{g(x)}\right| = +\infty$, we can say that $\lim \frac{f(x)}{g(x)} = \infty$ and we consider f to be an infinitesimal of lower order than g.

3.5. Determine whether the following implications are true or false:

$$\begin{aligned} a_n &\sim b_n \Rightarrow a_n - b_n, \to 0, \\ a_n &\sim b_n \Rightarrow e^{a_n} \sim e^{b_n}, \\ a_n &\sim b_n \Rightarrow (a_n)^a \sim (b_n)^a, \text{ for every } a \in \mathbb{R}. \end{aligned}$$

3.6. Compare the two sequences $a_n = n!$ and $b_n = n^n$.

3.7. Is it true that $\ln(1 + e^n)$ is an infinity of higher order than n, and that the difference $\ln(1 + e^n) - n$ diverges to $+\infty$?

3.8. Compute the limits

$$(a) \lim_{x \to -\infty} \sqrt[3]{x+1} e^x, \quad (b) \lim_{x \to 0} \left(x^2 - 3x\right) \ln x, \quad (c) \lim_{x \to 0^+} x e^{1/x}.$$

3.9. Compute the limits

$$(a) \lim_{x \to 1} \frac{\ln x}{x-1}, \quad (b) \lim_{x \to 0} \frac{1 - \cos x}{x^2}, \quad (c) \lim_{x \to 2} \frac{\sqrt{x} - \sqrt{2}}{x-2}.$$

3.10. Show that, as $x \to +\infty$,

$$\ln\left(3x^2 + 4x - 5\right) = 2 \ln x + \ln 3 + o(1).$$

3.11. Which of the following functions has the straight line $y = x$ as an oblique asymptote, when $x \to +\infty$?

$$f(x) = x + \ln x, \; g(x) = x + \sqrt[3]{x}, \; h(x) = x + \sin x, \; k(x) = x + \frac{\sin x}{x}.$$

4

Continuity

The concept of *continuity* is certainly one of the most important in Mathematics, both for theory and applications. By means of such a notion we can describe various kinds of phenomena showing "little sensitivity" to small variations of the relevant quantities. To clarify in a rigorous way this kind of behaviour we resort to limit operations. The development of the chapter is as follows.

• After a brief introduction we introduce the notion of continuity and we study the main properties of continuous functions. In particular:

— the Intermediate Value Theorem;

— Weierstrass's Theorem (about the existence of a global maximum and a global minimum).

4.1 An intuitive idea of continuity

We try now to illustrate what we mean by phenomena showing "little sensitivity" to small variations of the relevant quantities.

In order to help our intuition, think of driving a car and varying the pressure on the accelerator pedal *slightly*. In a normal situation, you will note a *slight* variation in the speed.

In Economics it is also easy to find situations of this kind. If in a production process there is a slight variation of the amount of the raw material to be processed, the physical volume of production will usually undergo a small variation as well. If we introduce a small variation of the availability of some goods for a consumer, the change in his satisfaction will generally be very modest. If we modify by a few Euro the income flow given by a big securities portfolio, the values of indexes estimating the performance of the manager vary in an insignificant way. We could produce

a large number of examples having a common structure which can be intuitively described in this way:

small variations of the causes produce small variations on the effects

or:

effects vary continuously with the causes producing them.

In the previous chapters we saw that, in the simplest cases, the *cause-effect* or *input-output* relation is described by a function $y = f(x)$ where x takes the role of the cause/*input* (in the previous examples: the pressure on the accelerator, the amount of the raw materials or consumer goods, the cash flow of a financial portfolio) and $f(x)$ plays the role of the effect/*output* (speed, production, satisfaction, performance). Thus all we have seen above can be described by the following scheme, which can be considered as a rough definition of the *continuity* of f at the point x, where E (the error) is extremely small if Δx is very close to zero.

$$x + \Delta x \longrightarrow \boxed{f} \longrightarrow f(x) + E.$$

Let us now try to compare the graphs in the figure.

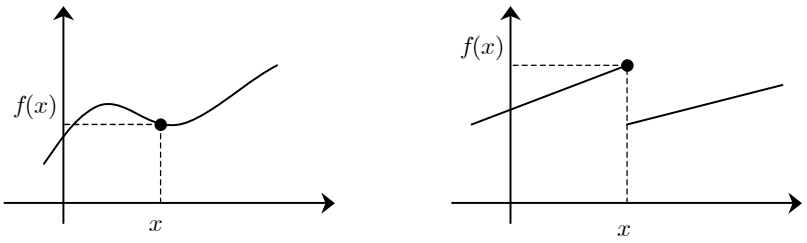


Figure 4.1.

In the first graph, if on the horizontal axis we move a bit from the point x, the values of f on the vertical axis remain close to the value $f(x)$: f is continuous at the point x. In the second one, increasing x a bit, a sudden variation (discontinuity) is produced on the values of $f(x)$: f is not continuous at x.

The importance of continuity is not limited to the description of the features of a law of dependence. Often, handling quantities that are changing in a *continuous* way is really necessary. When we study even the simplest mathematical models, we are faced with the solution of equations. Usually for equations such as

$$f(x) = 0$$

there are no explicit solutions (giving $x = \cdots$). Dealing with these equations, especially in view of concrete applications, constitutes a typical problem of *numerical analysis.* And there are more or less standard procedures, called *algorithms*, used

by widely available computer software, like *Maple, Mathematica, Mathcad* etc., in order to calculate approximate solutions.[1].

The use of codes for automatic calculation requires a certain caution and a test of the goodness of the approximation. Every operation with real numbers, obtained in an automatic way, is inexorably affected by some error; consider that even the use of the innocuous $\sqrt{2}$ determines an approximation error, simply because only a finite number of its (infinitely many) decimal digits may be calculated by a computer. The previously mentioned algorithms include long sequences of calculations with real numbers and thus it is indispensable to avoid that errors propagate and increase from one step to another. It is then necessary that *small errors in* x produce *small errors in the computation of* f, i.e. it is necessary for the function f we are dealing with to be *continuous.*

4.2 Continuous functions

The limit operation is the suitable tool for giving a precise meaning to the concept of continuity. A function f is continuous at a point x_0 if, *slightly* changing x_0, i.e. considering the point $x_0 + \Delta x$ which is very close to x_0, the corresponding value $f(x_0 + \Delta x)$ is *slightly* different from $f(x_0)$, i.e. with an error E which is very close to zero: the smaller $|\Delta x|$ is, the smaller is $|E|$. This can be expressed by saying that E tends to 0 if Δx tends to 0 and this is, in turn, equivalent to saying that $f(x_0 + \Delta x)$ tends to $f(x_0)$.

Definition 2.1. *Let f be defined in an interval $I \subseteq R$ and $x_0 \in I$. We say that f is* ***continuous*** *at x_0 if*

$$\boxed{\lim_{x \to x_0} f(x) = f(x_0)} \tag{4.1}$$

We say that f is *continuous in an interval* I if f is continuous at every point of I.

We can interpret (4.1) as the possibility of exchanging the operations of limit and function. Actually, if $\{x_n\}$ is a sequence tending to x_0 and f is continuous at that point, we have

$$\lim_{n \to +\infty} f(x_n) = f(x_0) = f\left(\lim_{n \to +\infty} x_n\right).$$

Then continuous functions are those which display a compatibility with the limit operation, so that *it is sufficient to calculate* f at x_0 in order to calculate the limit for x tending to x_0!

4.2.1 Continuity of elementary functions

Powers, exponentials, logarithms, *sine* and *cosine* functions are continuous at *every point of their natural domain.* Actually, we already noted in section 3.3 that, for every real number x_0 belonging to the natural domain of the considered function,

[1] The simplest of these procedures is introduced later, on page 113.

we have

$$\lim_{x\to x_0} x^\alpha = x_0^\alpha, \quad \lim_{x\to x_0} a^x = a^{x_0}, \quad \lim_{x\to x_0} \log_a x = \log_a x_0,$$
$$\lim_{x\to x_0} \sin x = \sin x_0, \quad \lim_{x\to x_0} \cos x = \cos x_0.$$

• (⇛ **Chapter 11**) *Decomposability.* Do we gain or do we lose if we interrupt an investment and then we invest again at the same conditions? Suppose we invest the amount $P = 20000$ Euro from time 0 to time t (years). We denote by $f(t)$ the total amount at time t for each unit of capital, so that at time t we receive $P \cdot f(t)$ Euro. After disinvesting, we reinvest the amount which we have just collected for s more years, at the same conditions. The final value we collect would be $P \cdot f(s) \cdot f(t)$ Euro.

If we had not interrupted the investment, at the end the final amount available would have been $P \cdot f(t+s)$. In order to understand whether we have taken advantage of the interruption or not, we just have to compare

$$f(s) \cdot f(t) \quad \text{with} \quad f(t+s),$$

so that, in general, the answer depends on the type of function f describing the accumulation law.

In conditions of simple interest with annual rate i, we have $f(t) = 1 + it$; in this case

$$\begin{aligned} f(s) \cdot f(t) &= (1+is)(1+it) = 1 + (s+t)i + i^2 st \\ f(t+s) &= 1 + (t+s)i \end{aligned}$$

whence, if $s > 0$, $t > 0$, since $i^2 st$ is positive, we have $f(s) \cdot f(t) > f(t+s)$ and the disinvestment is advantageous.

In conditions of compound interest we have $f(t) = (1+i)^t$ and then

$$f(s) \cdot f(t) = (1+i)^s (1+i)^t = (1+i)^{s+t} = f(s+t)$$

whence we deduce the equivalence between the two operations.

If for every positive s, t we obtain

$$f(s) \cdot f(t) = f(s+t) \tag{4.2}$$

financial mathematicians say that f is *decomposable.* If $f(t)$ is an exponential function, then it is decomposable. Conversely, if we limit ourselves to continuous functions, the only solutions of equation (4.2), which is called *Cauchy's functional equation,* are exponential functions; to be more precise:

If f is a (non constant) solution of (4.2)*, continuous in* $[0, +\infty)$*, then*

$$f(t) = a^t$$

for some positive a*.*

Working with continuous functions, we have found a simple and important characterization of the exponential function. Such a characterization is interesting not only from a mathematical point of view. This property, as well as other properties of the exponential function, is an important "technical" element in favour of the law of compound interest, which ought to be the only financial law in current use.

Continuity and algebraic operations

Combining continuous functions by means of algebraic operations, we again get continuous functions. This is an immediate consequence of the results in section 5.2 of Chapter 3.

Theorem 2.1. *Let f and g be continuous functions in an interval $I \subseteq \mathbb{R}$. Then $f+g$, $f \cdot g$ are continuous in I. If $g \neq 0$ in I, f/g is also continuous in I.*

Composite functions

The composition operation, also, does not destroy continuity.

Let $f : I \to J$ and $g : J \to \mathbb{R}$ where I, J are intervals, so that the composite function $g \circ f : I \to \mathbb{R}$ is well defined. If f is continuous in I and g is continuous in J then $g \circ f$ is continuous in I. More precisely

Theorem 2.2. *If f is continuous at $x_0 \in I$ and g is continuous at $y_0 = f(x_0)$ then $g \circ f$ is continuous at x_0.*

Proof. We have to prove that, if $\{x_n\}$ is a sequence converging to x_0, the sequence $\{g[f(x_n)]\}$ converges to $g[f(x_0)]$. Put $y_n = f(x_n)$. Continuity of f implies that $y_n \to y_0$, while continuity of g implies that $g(y_n) \to g(y_0)$, which is our thesis. □

The elementary functions considered in Chapter 2 are continuous at every point of their natural domain. As a consequence we deduce, for example, that functions like the following

$$f(x) = e^{-x^2}, \qquad g(x) = \ln(1+x^4), \qquad h(x) = \sin(e^{-x}+x)$$

are continuous in their natural domain.

4.2.2 Discontinuities

Requirement (4.1) is equivalent to the following three conditions: the two limits

$$\lim_{x \to x_0^+} f(x) \quad \text{and} \quad \lim_{x \to x_0^-} f(x)$$

— *exist* and are *finite*,
— they are *equal*,
— their common value is precisely $f(x_0)$.

If *at least* one of the requirements is not satisfied, then (4.1) is not true and we say that the function f is *discontinuous* at x_0 or that f has a *discontinuity* at x_0. In particular, we say that f displays a **jump discontinuity** at x_0 if the right-hand and left-hand limits at the point x_0 both exist, are both finite, but they are different from each other, that is if

$$\lim_{x \to x_0^+} f(x) = L \quad \text{and} \quad \lim_{x \to x_0^-} f(x) = L'$$

with $L \neq L'$. The quantity

$$J(x_0) = \lim_{x \to x_0^+} f(x) - \lim_{x \to x_0^-} f(x).$$

is called the *jump* of f at x_0.

• *Discounted delivery tariffs.* Imagine a forwarding agent offering the transport of goods for a payment which is a function of the weight x of the package to be delivered. The agent wants 1 Euro per kg for packages whose weight is less than 10 kg. From 10 kg onwards, the tariff applied for the whole package decreases to 0.9 Euro per kg. The transport cost $c(x)$ is then

$$c(x) = \begin{cases} x \text{ for } x < 10 \\ \\ 0.9x \text{ for } x \geq 10. \end{cases}$$

The graph of the tariff shows that at point 10 the function c is discontinuous because

$$\lim_{x \to 10^-} c(x) = 10 \quad \text{while} \quad \lim_{x \to 10^+} c(x) = c(10) = 9.$$

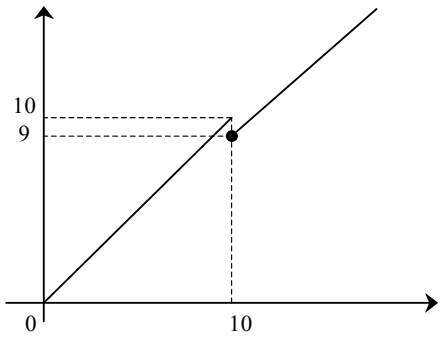


Figure 4.2. Jump discontinuity

In our example, it is cheaper to send a package weighing 10 kilos than one weighing 9.5 kilos!

An important example of functions that are continuous or display only jump discontinuities are monotonic functions.

Theorem 2.3. *If f is monotonic in the interval (a, b), then f has at most a countable number of points of discontinuity, which are all jump points.*

It may also happen that right-hand and left-hand limits are finite and equal, but they are different from the value of f:

$$\lim_{x \to x_0^+} f(x) = \lim_{x \to x_0^-} f(x) \neq f(x_0).$$

Discontinuities of this kind might be called *holes.* It is clear that we are dealing with discontinuities having a quite unnatural aspect: we could *eliminate* them by modifying the value of f at x_0, giving it the value of the two limits - and the discontinuity would disappear. In applications we hardly ever meet functions having this kind of behaviour, whereas their technical use is common in Probability and Statistics, where sometimes functions "highlighting" a certain number are needed. Whenever we want to give a statistical estimate of the value of an unknown parameter (for instance the mean income of a population) we can use a function describing the cost of the estimation error. The function of the error x

$$l(x) = \begin{cases} 0 \text{ if } x = 0 \\ \\ 1 \text{ if } x \neq 0 \end{cases}$$

is popular[2] without merit. For this function

$$\lim_{x \to 0} l(x) = 1 \neq l(0) = 0.$$

Right-hand and left-hand continuity

If a function is discontinuous at the point x_0, it may happen anyway that one of the two limits, left-hand or right-hand limit, is equal to $f(x_0)$. In this case we shall say, respectively, that f is *continuous from the right* or *from the left at* x_0.

The cost function of the forwarding agent we have just met is continuous from the right.

4.3 Properties of continuous functions

There are some properties, expressed in the following theorems, that make continuous functions particularly useful.

The statement of the first theorem, known as Bolzano's theorem (or the "zeros theorem") is intuitive in the light of the graph in Figure 3: if f is a continuous function in an interval $[a, b]$ and the values assumed at the endpoints a, b have opposite sign, its graph must cross the x-axis at least at one point. Each of these points is called a *zero of* f because at that point the value of the function is zero.

The catholic priest Bernard Bolzano (1781-1848) was the first to state that this "obvious" assertion needed a proof. From one of the possible proofs, which we reproduce here, we can obtain an algorithm for the computation of the solutions for an equation $f(x) = 0$, as we mentioned in the introduction.

[2] In statistical inference, the most popular method of estimation, called the method of the *greatest likelihood*, may be justified by the use of the considered "loss function". Actually, $l(x)$ should be the cost of an error in the estimation of the amount x.

Bolzano's Theorem

Theorem 3.1. *Let f be continuous in the interval $[a, b]$. If the values of f at the endpoints of the interval have opposite sign, that is if $f(a) \cdot f(b) < 0$, then f has at least one zero in (a, b), i.e. there is at least one point $z \in (a, b)$ such that*

$$f(z) = 0.$$

If f is strictly monotonic, then the zero is unique.

Proof. Suppose that $f(a) < 0$ and $f(b) > 0$.
We use a *dichotomy* procedure.
First step. We take the mid-point $c = (a+b)/2$ of the interval $[a, b]$, compute $f(c)$ and consider the three possibilities.

- If $f(c) = 0$ we have finished; therefore suppose that $f(c) \neq 0$.
- If $f(c) > 0$, in the interval $[a, c]$ the function f is continuous and takes values with opposite sign at the endpoints. Then put $a_1 = a$, $b_1 = c$.
- If $f(c) < 0$, in the interval $[c, b]$ the function f is continuous and takes values with opposite sign at the endpoints. Then put $a_1 = c$, $b_1 = b$.

Second step. Consider the mid-point $c_1 = (a_1 + b_1)/2$ of the interval $[a_1, b_1]$, compute $f(c_1)$ and consider the three possibilities, as we did before, $f(c_1) = 0$, $f(c_1) > 0$, $f(c_1) < 0$. If $f(c_1) = 0$ the theorem is proved; otherwise set $a_2 = a_1$, $b_2 = c_1$ if $f(c_1) > 0$ or $a_2 = c_1$, $b_2 = b_1$ if $f(c_1) < 0$.
Repeat then on $[a_2, b_2]$ the same operation applied to $[a_1, b_1]$ etc.

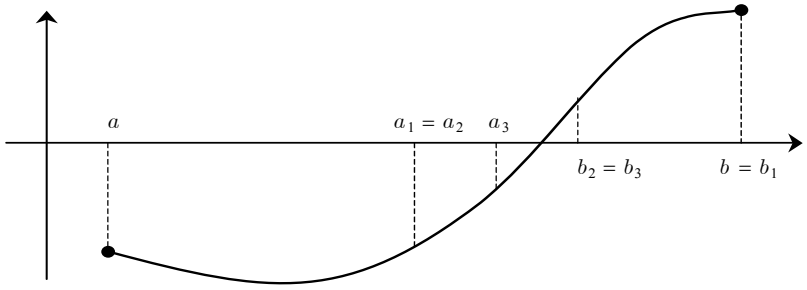


Figure 4.3. Bolzano's Theorem

Proceeding this way, two cases are possible:
(i) at one of the mid-points c_n the function f is zero. Then the dichotomy process stops and the proof is finished;
(ii) we construct a sequence of nested intervals

$$[a, b] \supset [a_1, b_1] \supset \cdots \supset [a_{n+1}, b_{n+1}] \supset \cdots$$

such that each interval is half the previous one. Thus

$$b_n - a_n = \frac{b-a}{2^n}$$

and this tends to zero for $n \to +\infty$. Moreover, the bisection strategy guarantees that the sequence of the left-hand endpoints $\{a_n\}$ is increasing, while the sequence of the right-hand endpoints $\{b_n\}$ is decreasing. Being monotonic and bounded, the sequences $\{a_n\}$ and $\{b_n\}$ converge and, what is more, they converge to the same limit, because $b_n - a_n \to 0$. Let us call this limit z :

$$\lim_{n\to+\infty} a_n = \lim_{n\to+\infty} b_n = z.$$

We show that $f(z) = 0$. To this purpose, we recall that the function f takes values with opposite sign at the endpoints of intervals $[a_n, b_n]$, that is

$$f(a_n) \cdot f(b_n) < 0. \tag{4.3}$$

If we take the limit in (4.3), using the theorem on the permanence of sign, we get

$$\lim_{n\to+\infty} f(a_n) \cdot f(b_n) \leq 0$$

whereas, if we use the continuity of f,

$$\lim_{n\to+\infty} f(a_n) = f\left(\lim_{n\to+\infty} a_n\right) = f(z), \quad \lim_{n\to+\infty} f(b_n) = f\left(\lim_{n\to+\infty} b_n\right) = f(z),$$

whence

$$[f(z)]^2 \leq 0.$$

Therefore we deduce that $f(z) = 0$. If f is also strictly monotonic, it cannot take the same value at two different points, so the zero is unique. □

The *dichotomy method* used in the above proof suggests a procedure for computing an approximation of the zeros of f. In fact if we stop the process after n steps, we can assume a_n (or b_n) as an approximate value of the zero[3] z with an error which is not greater than $b_n - a_n = (b-a)/2^n$.

We point out that if f has more than one zero, the method does not clarify which of them has been approximated; thus one should try to limit the procedure to intervals where there is only one zero, for example because the function is strictly monotonic there.

Sometimes the uniqueness of the solution may be determined by means of graphic considerations, like in the following example.

Example 3.1. Using a pocket calculator, let us try to compute the solution of the equation

$$f(x) = 2^{-x} - x = 0. \tag{4.4}$$

[3]If we take $(a_n + b_n)/2$ as an approximate value, we get nearer in many cases, but we lose information about the sign of the error (in excess or in defect).

The solution coincides with the abscissa of the intersection point of the graphs of $g(x) = 2^{-x}$ and $h(x) = x$. As we can see from the diagram, the solution is unique and it turns out to be between 0 and 1. Let us use the dichotomy method.

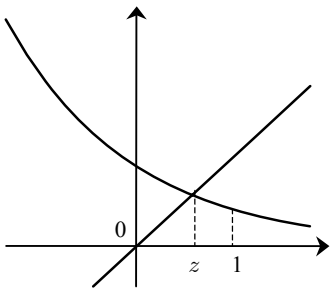


Figure 4.4.

We have $f(0) = 1 > 0$ and $f(1) = 2^{-1} - 1 = -0.5 < 0$. Subsequently:
$f(1/2) = 2^{-1/2} - 1/2 \simeq 0.20711 > 0 \Rightarrow$ we select the interval $[1/2, 1]$
$f(3/4) = 2^{-3/4} - 3/4 \simeq -0.1554 < 0 \Rightarrow$ we select the interval $[1/2, 3/4]$
$f(5/8) = 2^{-5/8} - 5/8 \simeq 0.02342 > 0 \Rightarrow$ we select the interval $[5/8, 3/4]$
$f(11/16) = 2^{-11/16} - 11/16 \simeq -6.6571 \times 10^{-2} < 0 \Rightarrow$ we select the interval $[5/8, 11/16]$
$f(21/32) = 2^{-21/32} - 21/32 \simeq -2.1725 \times 10^{-2} < 0 \Rightarrow$ we select the interval $[5/8, 21/32]$
$f(41/64) = 2^{-41/64} - 41/64 \simeq 8.001 \times 10^{-4} > 0$.

Stopping at the sixth iteration, we can state that the zero z is between $41/64 \simeq 0.64063$ and $21/32 \simeq 0.65625$

$$0.64063 < z < 0.65625$$

that is $z = 0.6...$. where the second digit is uncertain between 4 and 5.

Intermediate Value Theorem

A second property for continuous functions in an interval is called the *Intermediate Value Theorem* or *Darboux's Theorem*[4]. It can be expressed by saying that if f is a continuous function in an interval its range has no "holes". Precisely, if $f(x_1), f(x_2)$ are two elements of the range, then f takes all the values between $f(x_1)$ and $f(x_2)$. A discontinuous function may not fulfill this property, as shown by the second graph in Figure 5.

[4]Jean G. Darboux (1842-1917).

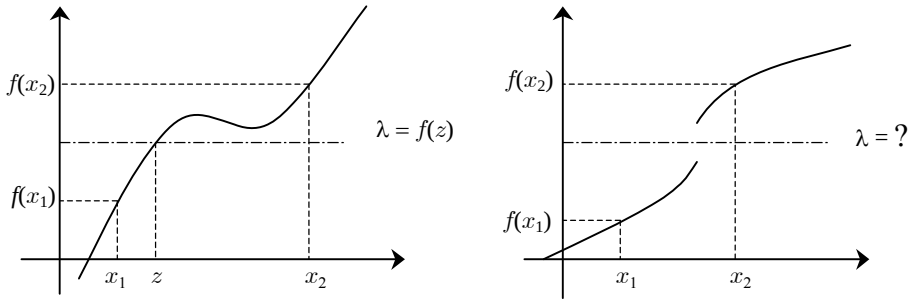


Figure 4.5. The intermediate value property

From the intermediate value theorem, theorem 3.1 follows immediately: if at the endpoints of an interval a (continuous) function f has values with opposite signs, then it takes all of the intermediate values and in particular the value 0. Conversely, from theorem 3.1 we can easily obtain the intermediate value property.

Theorem 3.2. *Let f be a continuous function in an interval $I \subseteq \mathbb{R}$ and let $x_1, x_2 \in I$. If $f(x_1) \neq f(x_2)$ and λ is a real number between $f(x_1)$ and $f(x_2)$, then there exists at least one point z between x_1 and x_2 such that $f(z) = \lambda$.*

Proof. With no loss of generality, let $x_1 < x_2$, $f(x_1) < f(x_2)$ and

$$f(x_1) < \lambda < f(x_2).$$

Then the function $g(x) = f(x) - \lambda$ is continuous in the interval $[x_1, x_2]$, and moreover

$$g(x_1) = f(x_1) - \lambda < 0, \quad g(x_2) = f(x_2) - \lambda > 0.$$

By theorem 3.1 there exists a point z in $[x_1, x_2]$ such that $g(z) = 0$, which is equivalent to $f(z) = \lambda$. □

Invertibility and continuity

At the beginning of this chapter we mentioned the importance of dealing with continuous functions when trying to find an approximate solution of a given equation. On the other hand, the problem of solving an equation is strictly connected with the problem of finding the inverse of a function. Indeed, any equation may be written in the form

$$f(x) = y$$

and for its solution we have to determine the value or the values of x having y as image. If f is invertible, for every y in the range of f there exists only one solution $x = g(y)$ and g is precisely the inverse function of f.

Now suppose that f is *invertible and continuous*. What can be said about the continuity of g? There are no problems for functions defined in an interval, for which the following theorem holds.

Theorem 3.3. *Let $f : (a, b) \to (c, d)$ be continuous and one-to-one. Then the inverse function g is continuous in (c, d).*

We end this paragraph with a remark. If f is continuous in (a, b) and one-to-one, then *it must be strictly monotonic*. Indeed, if there were three points $x_1 < x_2 < x_3$ such that, for instance $f(x_1) < f(x_2)$ and $f(x_2) > f(x_3)$, the intermediate value theorem assures us that between x_1 and x_2 and between x_2 and x_3 there must be two points where f takes the same value.

Note that we may find *non-strictly monotonic invertible* functions; if they are defined in an interval they cannot be continuous, of course! An example is shown in Figure 6.

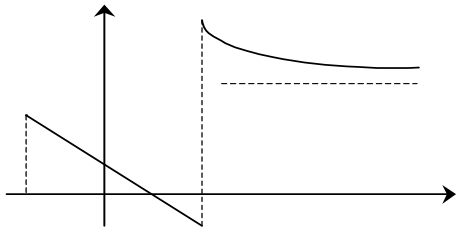


Figure 4.6. A non-monotonic invertible function

Maxima and minima. Weierstrass's Theorem

Continuous functions in closed and bounded intervals have global maxima and minima. This is a basic property in optimization problems. A large part of human activity is aimed at optimizing an objective: minimizing costs or losses, maximizing profits, saving energy, minimizing the time spent on an activity and so on. The study of these problems may be expensive from an analytical and computational point of view. Before investing resources in the search for a maximum (or minimum) it is worth checking its existence. The following theorem, due to K. Weierstrass (1815-1897), is therefore quite relevant, also because it is a prototype for more general results on the existence of optimal solutions.

Theorem 3.4. *If f is continuous in $[a, b]$ then it attains its maximum and its minimum value, that is there exist at least two points x_1 and x_2 in $[a, b]$, such that, for every x in $[a, b]$, we have*

$$m := f(x_1) \leq f(x) \leq f(x_2) =: M.$$

Therefore M is the maximum and m the minimum value of f. Remember that M, m belong to the range of f. Whereas the maximum or the minimum of f are unique, we may have many or even infinitely many points at which such values are attained.

Note that for a function f continuous in $[a, b]$ the Intermediate Value Theorem may now be stated as follows.

If f is continuous in $[a, b]$, the image of $[a, b]$ is the interval $[m, M]$.

We realize that this is a special property of continuous functions in an interval $[a, b]$, by examining the graphs in Figure 7.

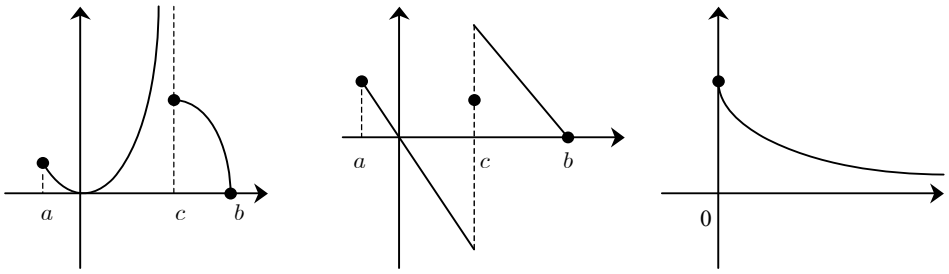


Figure 4.7. Existence/non-existence of maxima and minima

The first function is not even bounded, having a vertical left-hand asymptote at c. The second one is bounded but has neither maximum nor minimum. These functions are defined in a closed and bounded interval, but they are not continuous. The third one is continuous but it is defined in an unbounded interval and it has no minimum, even though it is bounded.

• (⇛ **Chapter 11**) *A leasing contract.* We talk about a leasing operation each time an individual or a firm, wishing to have some equipment available and not willing to acquire the property (or not being able to afford it), obtains the right to use it from a company. The beneficiary pays an established amount of money at fixed times to the leasing company and rents such equipment, which still remains the property of the company.[5] At the end of the contract, the equipment may be redeemed, so that it becomes the property of the individual or the firm that used it.

For example, let us consider a possible leasing contract for a car with value equal to 20000 Euro. The immediate down-payment is 20% of the value (that is 4000 Euro). Then, the customer has to pay 23 monthly leasing fees of 800 Euro each. Finally after 2 years and a payment of 200 Euro (the redemption price), she becomes the owner of the car. The total amount paid is

$$4000 + 23 \cdot 800 + 200 = 22\,600 > 20000.$$

Thus such payments include interests. But at what rate?

Denoting the cost of the car by A, the duration of the contract in months by n, the down-payment at the beginning of the contract by B, the sum (fee) paid at the end of each month by C, the redemption price by E, the *monthly compound interest rate* by x, we have a "model" of the *leasing* contract given by the formula

$$A = B + \sum_{k=1}^{n-1} \frac{C}{(1+x)^k} + \frac{E}{(1+x)^n}.$$

It is obtained by requiring the value A (cost of the car) to be equal to the sum of the down-payment, the discounted values of fees and the discounted redemption price.

[5] Usually the leasing company does not produce anything: it buys the equipment and then allows the applicant to use it.

In our example, we have

$$20000 = 4000 + \sum_{k=1}^{23} \frac{800}{(1+x)^k} + \frac{200}{(1+x)^{24}} \tag{4.5}$$

and the interest rate we are looking for is the solution of equation (4.5), which obviously cannot be solved algebraically. Let us rewrite (4.5) as

$$f(x) = 16000 - \sum_{k=1}^{23} \frac{800}{(1+x)^k} - \frac{200}{(1+x)^{24}} = 0. \tag{4.6}$$

It can be checked that $f(0.01) < 0$ and $f(0.02) > 0$ (that is $f(1\%) < 0$ and $f(2\%) > 0$) and since f is an increasing continuous function, equation (4.5) will have a unique solution between 0.01 and 0.02. Using numerical methods, we get

$$x \simeq 1.28\%.$$

4.4 Exercises

4.1. In a tax-system a personal income of amount x Euro is taxed at the mean rate

$$\alpha(x) := \begin{cases} 0 & \text{for } x \leq 10000 \\ 20\% & \text{for } 10000 < x \leq 30000 \\ 30\% & \text{for } x > 30000. \end{cases}$$

Denote by $f(x)$ the due tax. Draw a graph of the function $f(x)$ for $x \geq 0$. What happens at points $x = 10000$ and $x = 30000$?

4.2. Let the total revenue function r for a firm be defined as a function of the amount of production q:

$$r(q) = \begin{cases} q/2 & 0 \leq q < 100 \\ a \ln q + b & 100 \leq q < 1000 \\ 100 & q \geq 1000. \end{cases}$$

Determine for which values of the real parameters a and b the function r is continuous.

4.3. Find an interval including a solution of the equation

$$\ln x + x^2 = 0.$$

Then, by means of the bisection method, determine a solution within an error smaller than 1/100.

4.4. Decide whether the following statements are true or false.

(a) If a function defined in an interval $[a, b]$ is unbounded, then it has at least one discontinuity.

(b) A continuous function defined in an interval (a, b) surely attains its maximum and minimum value.

5

Differential Calculus and Optimization

Differential calculus (*Calculus*, for short) originates in the seventeenth century, from attempts to solve kinematic problems (such as the assessment of the velocity or the acceleration of a moving object) or geometric ones (such as the rectification of curves or the determination of the tangent line to a plane curve). Initially, this calculus develops on the basis of concepts which are not very rigorous, such as *Newton's fluxions* (interpretable as the *velocity of variation*) or *Leibnitz's infinitesimals*. In the period from the end of the seventeenth century to the first half of the nineteenth century, however, the growth of Calculus is impressive and allows to solve important problems, connected with celestial mechanics, hydrodynamics and wave propagation.

The foundation of Calculus, with the gradual elimination of these almost contradictory entities, such as infinitesimals and fluxions, starts with Cauchy (1820), who bases everything on the definition of the limit, and it consolidates definitively with Weierstrass (1872).

The achievement of a high level of rigour leads to a loss of the original immediacy and the resort to intuition is now less simple than it was at the outset. To limit the effect of this drawback, we have chosen to introduce the main concepts of Calculus with intuitive and geometric considerations. The development of the chapter is the following.

• We illustrate the concept of the *derivative* and we introduce the main rules of calculus; in particular:
— the derivatives of elementary functions,
— the algebra of derivatives,
— the derivative of composite and inverse functions.

• We illustrate the concept of the *differential*, also using simple experiments with a computer.

• We introduce the main tools for the optimization of one-variable functions. In particular:
— Fermat's theorem about stationary points,
— the mean value theorem (Lagrange's theorem),
— a test for monotonicity and a first test for the identification of stationary points.

• We analyse the problem of the approximation of a function with polynomials. This leads to:
— Taylor's formula,
— a second test for the identification of stationary points.

• We introduce a test for convexity/concavity and an identification test for inflection points.

5.1 Derivative and tangent line

Differential calculus is the most appropriate tool for solving optimization problems, which are obviously significant in Economics. In the simplest situations we are interested in finding the values which maximize or minimize a function, which is the target we want to optimize; typical examples are the maximization of a consumer's utility or the minimization of a production cost. To solve problems of this type, we must analyse how the target function varies, according to the fluctuations of the variable (or the variables) on which it depends. We have already met a similar situation when dealing with the concept of continuity, but here a deeper analysis is needed: we need a study of the changes of the target function which allows us to separate the "relevant" part from the "irrelevant" part. For the moment we deal with targets which are represented by one-variable functions, following a route which will bring us in a natural way to the concepts of velocity, tangent lines, *derivatives* and *differentials*.

In particular, this latter concept concerns the possibility of well approximating the graph of a function "locally", that is near a point of its graph, with a line whose slope is precisely the derivative of the function at that point.

The two concepts are therefore so closely bound together that, for one-variable functions (but only for these functions), the existence of *derivatives* and *differentials* are equivalent properties.

Let f be defined on the interval (a, b) and let $x_0, x \in (a, b)$. We move along the graph of f from the point $P_0 = (x_0, y_0)$ to the point $P = (x, y)$; we have

$$y_0 = f(x_0), \qquad y = f(x).$$

By subtracting, we obtain the vertical increment $\Delta f = f(x) - f(x_0)$, which corresponds to a horizontal increment $x - x_0$. If we divide the vertical increment by $x - x_0$, we obtain

$$\frac{\text{vertical increment}}{\text{horizontal increment}} = \frac{\Delta f}{\Delta x} = \frac{f(x) - f(x_0)}{x - x_0}. \tag{5.1}$$

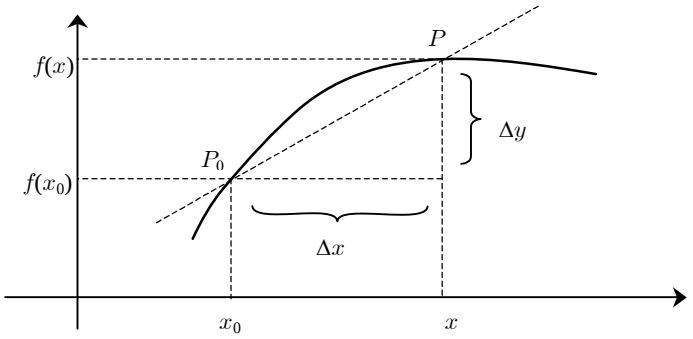


Figure 5.1. The difference quotient is the slope of the secant line PP_0

Definition 1.1. *The ratio* (5.1) *is called the* ***difference quotient*** *of* f *at the point* x_0, *relative to the increment* $x - x_0$.

The difference quotient represents the *average velocity*, or the *average rate of variation* of f relative to the path from x_0 to x. Geometrically, it is the *slope* of the secant line passing through the points P_0 and P.

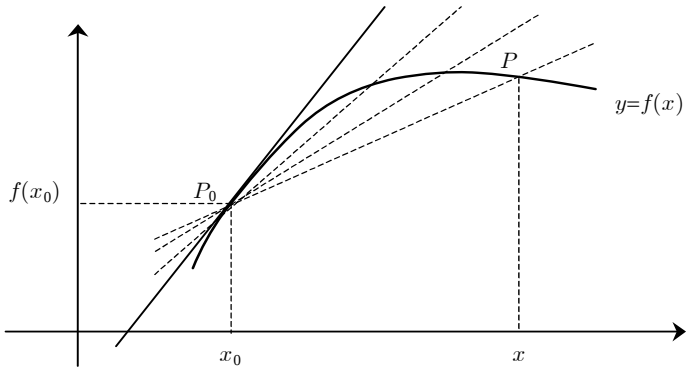


Figure 5.2. The secant line PP_0 tends to the tangent line at P_0

We now pass to the limit for $x \to x_0$ in (5.1) and we suppose that the limit exists: what happens? The point P moves along the graph of f towards P_0, which remains fixed (Figure 2). Consequently the line through P_0 and P, a secant line to the graph of f at P_0 and P, varies its slope and tends to a limit line. This line is called the *tangent line to the graph of* f at the point P_0; its slope will be called the *derivative of* f at the point x_0. Precisely, we have:

Definition 1.2. *Let* $f : (a, b) \to \mathbb{R}$ *and let* $x_0 \in (a, b)$; f *has a derivative at* x_0 *if the* ***finite*** *number*

$$\boxed{f'(x_0) := \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0}} \tag{5.2}$$

exists; this number is called the ***first derivative*** *(or simply the* ***derivative****) of* f *at* x_0.

The symbol $f'(x_0)$ (introduced by the Turinese Giuseppe Luigi Lagrange (1736-1813)) is read "f prime of x_0". Recalling that the difference quotient may be interpreted as the average rate of variation of f in the interval with extremes x_0 and x, it follows that the derivative may be interpreted as the *instantaneous rate of variation.*

Definition 1.3. *The line of equation*

$$\boxed{y = f(x_0) + f'(x_0)(x - x_0)}$$

is called the ***tangent line*** *to the graph of* f *at the point* $(x_0, f(x_0))$.

We consider two elementary examples immediately.

• *Affine linear functions* (*polynomials of degree one*). We consider a generic straight line of equation

$$y = mx + q$$

where m represents its slope. If we move from one of its points (x_0, y_0) to another point (x, y), we find

$$y - y_0 = mx - mx_0 = m(x - x_0) \tag{5.3}$$

and consequently

$$\frac{\Delta y}{\Delta x} = \frac{y - y_0}{x - x_0} = m.$$

The difference quotient $\Delta y/\Delta x$ does not depend on the point x_0 we choose at the beginning: it is always m. It follows that the derivative of an affine linear function at every point x_0 is m. In particular, *the derivative of a constant function is zero.*

• *Quadratic functions* (*polynomials of degree two*). We consider the function

$$f(x) = x^2,$$

whose graph is a parabola with vertex at the origin, symmetric with respect to the y-axis. Proceeding as above, we have

$$\Delta f = f(x) - f(x_0) = x^2 - x_0^2$$

and observing that $x^2 - x_0^2 = (x - x_0)(x + x_0)$ we find

$$\frac{\Delta f}{\Delta x} = \frac{x^2 - x_0^2}{x - x_0} = x + x_0 = (x - x_0) + 2x_0. \tag{5.4}$$

Differently from the linear case, the ratio $\Delta f/\Delta x$ is no longer a constant: it depends on the position of the points x and x_0. It represents the *slope* of the line passing through $P_0 = (x_0, y_0)$ and $P = (x, y)$. When $x \to x_0$,

$$\frac{\Delta f}{\Delta x} = (x - x_0) + 2x_0 \to 2x_0$$

while the point P moves along the parabola towards P_0 (Figure 3) and the secant line at P_0 and P tends to the tangent line to the parabola at the point P_0, whose equation is

$$y = x_0^2 + 2x_0(x - x_0). \tag{5.5}$$

The number $2x_0$, the slope of the tangent, is the *derivative of the function* $f(x) = x^2$ at the point x_0.

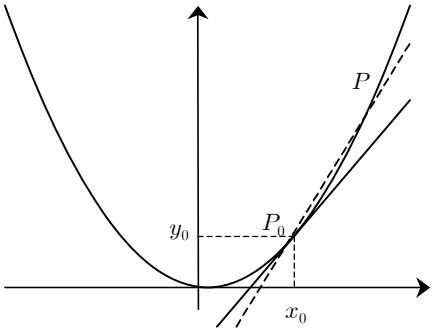


Figure 5.3. Tangent line to the parabola at P_0

Intuitively, if a function has a derivative it must be "smooth", that is its graph has no sharp corners. We should however say that in some statistical or economical applications we may sometimes run into functions with some "special" points. A typical function which is "non smooth" at $x_0 = 0$ is $f(x) = |x|$. Its graph has a sharp corner at the origin.

Remark (*Other notations*). Setting $h = x - x_0$, the definition of derivative may be rewritten in the form

$$\boxed{f'(x_0) := \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}} \tag{5.6}$$

which is more convenient for computations (obviously, if the limit exists and is finite).

5.1.1 Derivatives and continuity. Right and left derivatives

An immediate consequence is that a function having derivative cannot have jumps or other sorts of discontinuity. Indeed:

Theorem 1.1. *If f has a derivative at x_0 then f is continuous at x_0.*

Proof. We write

$$f(x_0 + h) - f(x_0) = \frac{f(x_0 + h) - f(x_0)}{h} h.$$

Passing to the limit for $h \to 0$, we obtain $f(x_0 + h) - f(x_0) \to f'(x_0) \cdot 0 = 0$, which is the continuity of f at x_0. □

The implication stated in theorem 1.1. cannot be reversed, since there exist functions which are *continuous but have no derivative*. A classical example is the function $f(x) = |x|$. It is a continuous function at the origin but it has no derivative at that point. Indeed, its difference quotient is

$$\frac{f(0+h)-f(0)}{h} = \frac{|h|-|0|}{h} = \begin{cases} 1 & \text{if } h > 0 \\ -1 & \text{if } h < 0. \end{cases}$$

Therefore, there exist both the right and left limits of the difference quotient (respectively equal to 1 and -1) but they are different; thus the derivative of f at $x_0 = 0$ does not exist.

The example we have just considered shows that the right and left limits of the difference quotient for $h \to 0$ may exist while the limit (5.6), which defines the derivative of f at the point x_0, may not exist. It may also be necessary to calculate increments of a function around a point x_0 which is an extreme point of an interval. In this case, indeed, only positive or negative increments of the variable x make sense, in order to remain within the domain of the function. Proceeding in this way, we are lead to the concepts of a right and left derivative.

Definition 1.4. *The **right** and **left derivatives** of f at x_0 are the limits (if they exist and are finite)*

$$f'_+(x_0) := \lim_{h \to 0^+} \frac{f(x_0+h)-f(x_0)}{h}, \qquad f'_-(x_0) := \lim_{h \to 0^-} \frac{f(x_0+h)-f(x_0)}{h}.$$

In the case $f(x) = |x|$, at $x = 0$ we have $f'_+(0) = 1$ and $f'_-(0) = -1$, and the graph has a sharp corner at the origin. To be precise, *when at a point x_0 both the right and left derivative of f exist but they are different, we say that the point is a **corner***.

5.1.2 Interpretations of the derivative

• *Kinetics*. We consider an object in motion along a line. We introduce a system of coordinates for the line, so that we can identify the position of the object at every instant. Let $s = s(t)$ be the function which gives, at every instant, the position of the object in motion (time-position law):

$$s(t) = \text{position at time } t.$$

If $s(t_0)$ is the position of the object at the time t_0 the difference quotient

$$\frac{s(t)-s(t_0)}{t-t_0}$$

defines the *average velocity* in the time interval (t_0, t). When $t \to t_0$, we obtain the *instantaneous velocity* of the object at the time t_0. Therefore:

$$s'(t_0) = \text{instantaneous velocity at } t_0.$$

• *Economics.* We consider a factory which has to sustain a total cost C in order to produce a certain quantity q of its product. We call $C = C(q)$ the cost as a function of the produced quantity q. If the factory shifts from a certain amount of production q_0 to another, q, it is useful to estimate the ratio between the increment of the production cost and the increment of the quantity which is produced. Such a ratio

$$\frac{C(q) - C(q_0)}{q - q_0}$$

is called the *average rate of cost variation* (relative to the variation from q_0 to q). The variation of C relative to an infinitesimal variation of q, that is the derivative $C'(q_0)$, is called the *marginal cost* at q_0.

There are other possible meanings of the derivative, for example

— if $v(t)$ is the velocity of a point which moves along a straight trajectory (a line), $v'(t)$ represents the *acceleration*;

— if $r(q)$ is the revenue as a function of the produced quantity q, the derivative $r'(q)$ is the *marginal revenue*;

— if $p(t)$ is the price of one kind of goods at time t, for example the price of a stock in a computerized Stock Exchange, $p'(t)$ may be regarded as the *instantaneous velocity of price variation.*

5.2 Elementary formulae

In order to compute the derivatives of functions, we need to master some rules . In the case of long and boring computations, we may however use a computer and a software (such as *Maple, Mathematica, Mathcad,...*) which is able to perform all the computations of symbolic calculus.

As we have seen in the previous section, for the function $f(x) = x^2$ we can define the derivative at every point x and this derivative is equal to $2x$. Together with the function $f(x) = x^2$, we have thus defined the function:

$$x \mapsto 2x.$$

Generally speaking, if f is defined in an interval (a, b) and has a derivative at every point x of (a, b), to every point x we can associate the derivative of f at x:

$$\boxed{x \mapsto f'(x)}$$

constructing in this way a function f', also defined on (a, b), which we call the *first derivative* or simply the *derivative* of f.

The notation f' for the derivative is due to Lagrange, as we have already mentioned. Other notations for the derivative of $y = f(x)$ are $\mathrm{D}f$ (Cauchy's notation),

$\dot{f}$ (Newton's notation[1]) and

$$\frac{\mathrm{d}f}{\mathrm{d}x} \quad \text{or} \quad \frac{\mathrm{d}y}{\mathrm{d}x} \qquad \text{(Leibniz's notation)}.$$

To give another example, we compute the derivative of the *power function*

$$f(x) = x^n \qquad (n \text{ natural} \geq 1)$$

at a generic point x. Denoting by h the increment of the independent variable x, we have

$$\Delta f = f(x+h) - f(x) = (x+h)^n - x^n.$$

We try to write the increment Δf as a sum of a *linear term in* h and of other terms which contain only powers of h whose exponent is larger than 1 and are therefore negligible when $|h|$ is small enough. To this purpose, we develop $(x+h)^n$ using Newton's binomial formula[2]; we obtain:

$$(x+h)^n = x^n + nx^{n-1}h + \text{powers of } h \text{ with exponent} \geq 2,$$

therefore

$$\frac{\Delta f}{h} = nx^{n-1} + \ o(1) \ \ \to nx^{n-1}$$

if $h \to 0$. In conclusion, the function $f(x) = x^n$ has derivative and we obtain

$$\boxed{f'(x) = nx^{n-1}.}$$

Derivatives of some elementary functions

The formula we have just found extends to the case in which the exponent is any real number α. Such a function is defined, in general, only for $x > 0$ and has a derivative there at every point. We have:

$$f(x) = x^\alpha, \quad f'(x) = \alpha x^{\alpha-1} \ .$$

In particular, we have:

— The function $f(x) = \dfrac{1}{x} = x^{-1}$ has a derivative (at every point) in its domain and its derivative is $f'(x) = (-1)\,x^{-1-1} = -\dfrac{1}{x^2}$.

— The function $f(x) = \sqrt{x} = x^{\frac{1}{2}}$, defined in $[0, +\infty)$, has a derivative in $(0, +\infty)$ and its derivative is $f'(x) = \dfrac{1}{2}x^{\frac{1}{2}-1} = \dfrac{1}{2\sqrt{x}}$.

---

[1] Newton's notation is often used in Economics, as well as in Physics, when the independent variable is time.

[2] Chapter 1, theorem 8.4.

— The function $f(x) = \sqrt[3]{x} = x^{\frac{1}{3}}$, defined in $\mathbb{R}$, has a derivative for $x \neq 0$ and its derivative is $f'(x) = \dfrac{1}{3}x^{\frac{1}{3}-1} = \dfrac{1}{3\sqrt[3]{x^2}}$.

If we look again at the graphs of power functions, we see that they have a vertical tangent line at $x = 0$, with infinite slope, if $0 < \alpha < 1$. These functions therefore do not have a derivative at $x = 0$. If instead $\alpha > 1$, their graphs pass through the origin with slope equal to zero (and therefore their right derivatives are equal to zero).

We now verify this. We start by computing the right limit of the difference quotient at $x = 0$, for $\alpha > 0$. We have:

$$\lim_{h\to 0^+} \frac{f(h) - f(0)}{h} = \lim_{h\to 0^+} \frac{h^\alpha}{h} = \begin{cases} +\infty & 0 < \alpha < 1 \\ 1 & \alpha = 1 \\ 0 & \alpha > 1. \end{cases}$$

Therefore, for power functions $f'_+(0)$ *exists only if* $\alpha \geq 1$. The computation of the left limit (if f is defined in $\mathbb{R}$) is analogous.

The exponential, logarithmic and trigonometric functions have a derivative at every point of their domain. We list the derivatives of the most common functions in the following table.

$f(x)$	k	x	x^α	e^x	$\ln x$	$\sin x$	$\cos x$
$f'(x)$	0	1	$\alpha x^{\alpha-1}$	e^x	$1/x$	$\cos x$	$-\sin x$

We have already dealt with the first three cases; let us check the remaining ones.

• We compute the derivative of the function $f(x) = e^x$. We have

$$\lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{e^{x+h} - e^x}{h}.$$

Collecting e^x and recalling the example 5.15 of Chapter 3, we find

$$\lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = e^x \lim_{h\to 0} \frac{e^h - 1}{h} = e^x.$$

For the function $g(x) = a^x$, $a > 0$, we have $g'(x) = a^x \ln a$. This result can be obtained by writing $g(x) = e^{x \ln a}$ and computing its derivative using the chain rule, which we shall see in section 4.

• For the derivative of the function $f(x) = \ln x$, we proceed analogously. We have:

$$\lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{\ln(x+h) - \ln x}{h}.$$

Applying the properties of the logarithmic function and recalling the example 5.14 of Chapter 3, we find:

$$\lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{\ln(1+h/x)}{h} = \frac{1}{x} \lim_{h\to 0} \frac{\ln(1+h/x)}{h/x} = \frac{1}{x}.$$

For the function $g(x) = \log_a x$, $a > 0$, $a \neq 1$, we have $g'(x) = \dfrac{1}{x \ln a}$. This result can be obtained by simply writing $g(x) = \dfrac{\ln x}{\ln a}$.

• The calculation of the derivatives of the trigonometric functions requires the knowledge of the limits computed in the examples 5.11 and 6.7 of Chapter 3. If $f(x) = \sin x$, we have:

$$\lim_{h\to 0} \frac{f(x+h) - f(x)}{h} = \lim_{h\to 0} \frac{\sin(x+h) - \sin x}{h}.$$

Applying the addition formula for the sine function, we find

$$\begin{aligned} \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} &= \lim_{h\to 0} \frac{\sin x \cos h + \cos x \sin h - \sin x}{h} = \\ &= \lim_{h\to 0} \left(\frac{\cos h - 1}{h} \sin x + \frac{\sin h}{h} \cos x \right) = \\ &= 0 \cdot \sin x + 1 \cdot \cos x = \cos x. \end{aligned}$$

The calculation of the derivative of the function $\cos x$ is analogous (and it is left to the reader).

5.3 Algebra of derivatives

Functions can be combined together using the common algebraic operations, obtaining this way the sum function, the product function, the quotient function, etc. We consider a factory which produces goods with a process consisting of two phases, with different functions for the total and marginal cost in each phase. The total cost of the process is described by the sum of the two functions of total cost and the marginal cost of the whole production process turns out to be the sum of the two marginal costs. This is a particular case of the rule: *the derivative of the sum of functions is the sum of the derivatives of each addend.*

Still considering production costs, if $C = C(q)$ represents, as usual, the cost required to produce the quantity q, the quotient $C_{\mathrm{A}}(q) = C(q)/q$ is the *average cost for a unit of product (or average unitary cost)* and it is economically relevant to calculate its derivative. We have to compute the derivative of the quotient of $C(q)$ by q; we will see that this is *not* the quotient of the derivatives. Analogously, we will see that the derivative of the product of functions is *not* the product of the derivatives.

Our goal is to describe the simple rules which allow us to calculate the derivative of a function obtained by combining other functions using the sum, product and quotient, if we know the derivatives of the component functions.

Linearity

We consider two functions f and g, both having derivatives at a point x, and their sum function $f+g$. If, for $x \to x_0$, we have

$$\frac{f(x)-f(x_0)}{x-x_0} \to f'(x_0) \quad \text{and} \quad \frac{g(x)-g(x_0)}{x-x_0} \to g'(x_0) \tag{5.7}$$

by conveniently summing and collecting terms we obtain

$$\frac{[f(x)+g(x)]-[f(x_0)+g(x_0)]}{x-x_0} \to f'(x_0)+g'(x_0) \tag{5.8}$$

Equation (5.8) reveals that the sum function has a derivative, and this is the sum of the derivatives. We summarize all this in the formula

$$\boxed{(f+g)' = f' + g'} \tag{5.9}$$

If we multiply the first of the equations (5.7) by a real number k, we obtain

$$\frac{kf(x)-kf(x_0)}{x-x_0} \quad \to \quad kf'(x_0)$$

from which we deduce the formula

$$\boxed{(kf)' = kf'} \tag{5.10}$$

The validity of formulae (5.9) and (5.10) shows that derivation may be considered as a *linear* operation which, given a function, produces another function.

Example 3.1. We calculate the derivative of the function $f(x) = 2x + 4x^3$. By recursively applying formulae (5.9) and (5.10), we have

$$\mathrm{D}f(x) = \mathrm{D}(2x) + \mathrm{D}(4x^3) = 2\mathrm{D}(x) + 4\mathrm{D}(x^3) = 2 + 4(3x^2) = 2 + 12x^2$$

(we have used Cauchy's notation here).

Product and quotient

We have already remarked that the formula for the derivative of a product is not as immediate as in the case of the sum. To find the correct result, we start with the difference quotient

$$\frac{f(x)\,g(x) - f(x_0)\,g(x_0)}{x-x_0},$$

we subtract and add $f(x_0)\,g(x)$ in the numerator:

$$= \frac{f(x)\,g(x) - f(x_0)\,g(x) + f(x_0)\,g(x) - f(x_0)\,g(x_0)}{x-x_0} =$$

we decompose:

$$= \frac{f(x)-f(x_0)}{x-x_0} g(x) + f(x_0)\,\frac{g(x)-g(x_0)}{x-x_0}$$

and we pass to the limit for $x \to x_0$:

$$\frac{f(x)\,g(x) - f(x_0)\,g(x_0)}{x - x_0} \to f'(x_0)\,g(x_0) + f(x_0)\,g'(x_0)\,.$$

We finally get:

$$\boxed{(f \cdot g)' = f'g + fg'} \tag{5.11}$$

Example 3.2. We calculate the derivative of the function $f(x) = x^2 e^x$. Applying (5.11), we have

$$\mathrm{D}f(x) = \mathrm{D}(x^2)e^x + x^2\mathrm{D}(e^x) = 2xe^x + x^2e^x = xe^x\,(2 + x)\,.$$

In the case of the quotient we limit ourselves, for the moment, to giving the formula, and we will prove it later on: if, besides the usual assumptions, g is not zero, then the quotient f/g has a derivative and the following formula holds:

$$\boxed{\left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}} \tag{5.12}$$

Example 3.3. We calculate the derivative of the function

$$f(x) = \frac{\ln x}{x} \quad \text{and} \quad g(x) = \tan x = \frac{\sin x}{\cos x}.$$

For the function f, applying (5.12) we have:

$$\mathrm{D}f(x) = \frac{\mathrm{D}(\ln x) \cdot x - \ln x \cdot \mathrm{D}(x)}{x^2} = \frac{1 - \ln x}{x^2};$$

for the function g, we have

$$\begin{aligned} \mathrm{D}g(x) &= \frac{\mathrm{D}(\sin x) \cdot \cos x - \sin x \cdot \mathrm{D}(\cos x)}{(\cos x)^2} = \frac{(\cos x)^2 + (\sin x)^2}{(\cos x)^2} = \\ &= \frac{1}{(\cos x)^2} = 1 + (\tan x)^2\,. \end{aligned}$$

The derivation rules for sum and product can be generalized to any number of addends or factors. If $f_1, f_2, \ldots, f_n$ are functions having a derivative, then

$$\begin{aligned} \left(\sum_{s=1}^{n} f_s\right)' &= \sum_{s=1}^{n} f_s' \\ \left(\prod_{s=1}^{n} f_s\right)' &= {f_1}' f_2 \cdots f_n + f_1 {f_2}' \cdots f_n + \cdots + f_1 f_2 \cdots f_{n-1} {f_n}'. \end{aligned}$$

Example 3.4. We compute the derivative of a generic polynomial. We have

$$\mathrm{D}\left(\sum_{s=0}^{n} a_s x^s\right) = \sum_{s=0}^{n} \mathrm{D}(a_s x^s) = \sum_{s=1}^{n} s a_s x^{s-1}.$$

Example 3.5. We compute the derivative of the function $f(x) = \ln x\sqrt{x}e^x$. We have

$$f'(x) = \frac{1}{x}\sqrt{x}e^x + \ln x\frac{1}{2\sqrt{x}}e^x + \ln x\sqrt{x}e^x.$$

Example 3.6. The costs $C = C(q)$ sustained by a factory to produce an amount q of goods can be subdivided into *fixed costs* f and *variable costs* $V(q)$. We assume that the *unitary* variable costs $v(q)$ grow (linearly) with q according to the formula

$$v(q) = aq + b \quad \text{with } a > 0,\ b \geq 0.$$

Since $V(q) = qv(q)$, we find

$$C(q) = f + qv(q) = f + q(aq + b) = aq^2 + bq + f.$$

In this case we have

$$C'(q) = 2aq + b.$$

The variation of costs determined by changing the production amount from the initial quantity q_0 to the quantity q is, for $|q - q_0|$ sufficiently small,

$$C(q) - C(q_0) \simeq (2aq + b)(q - q_0).$$

This relation tells us that the cost due to the supplementary production amount $q - q_0$ (this amount can be measured, for example, in kg) is approximately proportional to the amount $q - q_0$ itself. Therefore every supplementary unitary amount (for example every extra kg) which is produced costs around $2aq_0 + b$ units of price (for example Euro). This means that producing one more kg implies different supplementary costs according to the current amount produced q_0, say 2000 or 3000 kg, to which this increment adds. In the case we have considered, since one supplementary kg costs more and more as the amount produced grows larger and larger, we say that we have a *Diseconomy of Scale* (for the variable costs).

5.4 Composite functions and inverse functions

5.4.1 The derivative of a composite function

A factory fixes the selling price x of some goods. The factory knows the so-called demand function, that is the physical amount $q = q(x)$ of goods which will be demanded if the price is x. The factory also knows the cost function $C = C(q)$, that is the production cost of q units of goods.

The choice of the price x determines, via $q = q(x)$, the amount of the production costs $C[q(x)]$, which is the composite function of $q = q(x)$ and $C = C(q)$. Is it possible to obtain the derivative of the composite function from the derivatives of q and C? The answer is incredibly simple: "Yes, it is their product!"

This is a consequence of the rule for the derivative of composite functions, a rule which we now state and prove. Let f and g be two composable functions, that is if

x belongs to the domain of g then its image $y = g(x)$ belongs to the domain of f. In this case, it makes sense to speak of the composite function $f \circ g$ or, more explicitly, of the function $z = f[g(x)]$. If f and g have derivatives, we can express the derivative of the composite function by means of the derivatives of its components. Precisely:

Theorem 4.1. *If g has a derivative at x and f has a derivative at $g(x)$, then $f \circ g$ has a derivative at x and the following formula holds*

$$\boxed{(f \circ g)'(x) = f'[g(x)] \cdot g'(x).} \tag{5.13}$$

Formula (5.13) is also called the *chain rule*.

This becomes totally clear using Leibniz's notation. Indeed, setting $z = f(y)$ and $y = g(x)$, we have

$$(f \circ g)' = \frac{\mathrm{d}z}{\mathrm{d}x}, \quad f' = \frac{\mathrm{d}z}{\mathrm{d}y}, \quad g' = \frac{\mathrm{d}y}{\mathrm{d}x}$$

and we can therefore write (5.13) in the following form

$$\frac{\mathrm{d}z}{\mathrm{d}x} = \frac{\mathrm{d}z}{\mathrm{d}y} \cdot \frac{\mathrm{d}y}{\mathrm{d}x} \qquad \text{(as if the two symbols d}y \text{ cancel out).} \tag{5.14}$$

Formula (5.14) expresses the fact that the variation rate of z with respect to x is the product of the "intermediate" variation rates of z with respect to y and of y with respect to x.

Proof. Since f has derivative at $y = g(x)$, we have

$$f(y+k) - f(y) = f'(y)\,k + o(k), \quad \text{for } k \to 0 \tag{5.15}$$

Now, we can write $o(k)$ in the form $k\varepsilon(k)$ with $\varepsilon(k) \to 0$ if $k \to 0$ and $\varepsilon(0) = 0$. By choosing $k = g(x+h) - g(x)$, since g has a derivative at x, for $h \to 0$ we have

$$k \to 0 \qquad \text{and} \qquad \frac{k}{h} \to g'(x).$$

Consequently, if $h \to 0$ also $\varepsilon(k) \to 0$. Dividing (5.15) by h, since $y + k = g(x+h)$, we have

$$\frac{f[g(x+h)] - f[g(x)]}{h} = \frac{k}{h}\left(f'[g(x)] + \varepsilon(k)\right)$$

and passing to the limit for $h \to 0$ we obtain (5.13). □

Examples

4.1. We calculate the derivative of $h(x) = (\ln x)^4$. This function is obtained as the composition of the two functions $z = f(y) = y^4$ and $y = g(x) = \ln x$.

Since $f'(y) = 4y^3$ and therefore $f'(\ln x) = 4(\ln x)^3$, and since $g'(x) = \dfrac{1}{x}$, using (5.13) we have

$$h'(x) = 4(\ln x)^3 \frac{1}{x}.$$

Using the formula (5.14) we would obtain

$$\frac{\mathrm{d}h}{\mathrm{d}x} = \frac{\mathrm{d}f}{\mathrm{d}y} \cdot \frac{\mathrm{d}y}{\mathrm{d}x} = 4y^3 \frac{1}{x} = 4 (\ln x)^3 \frac{1}{x}.$$

4.2. The derivative of the power functions $f(x) = x^\alpha$, for $x > 0$, can be obtained from the exponential function via the chain rule. Indeed, for $x > 0$, we have

$$f(x) = x^\alpha = e^{\alpha \ln x},$$

implying that

$$f'(x) = e^{\alpha \ln x} \cdot \alpha \cdot \frac{1}{x} = \alpha x^{\alpha - 1}.$$

In a similar way we obtain the derivative for $x < 0$ (for the power functions which are defined also for negative values of x).

4.3. For every $x \neq 0$ we have

$$\mathrm{D} (\ln |x|) = \frac{1}{x}.$$

The formula, already known for $x > 0$, can also be proved for $x < 0$, by applying the chain rule

$$\mathrm{D} [\ln (-x)] = \frac{1}{-x} \cdot (-1) = \frac{1}{x}.$$

• *Income tax*. Let x be the number of annual working hours of a worker, and let $z = g(x)$ be the income earned without applying any income tax. Let, finally, $y = f(z)$ be the worker's income after income tax. That income $y = f[g(x)]$ is therefore a composite function which depends on the number of annual working hours. Let us now increase the number of working hours: while the income before tax *increases faster and faster* (the overtime hours are better paid than the normal working hours), the part of income remaining after tax *always increases, but slower and slower*, because of the progressive taxation on income (the tax rate is higher on larger incomes). It is therefore a serious problem to understand how the income after tax (that which remains for the worker) varies depending on x. The variation rate is

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}y}{\mathrm{d}z} \frac{\mathrm{d}z}{\mathrm{d}x} = f' [g(x)] g'(x).$$

This equality shows that the variation rate of income after tax with respect to the working hours is the product of two derivatives. For large numbers of working hours $g'(x)$ will be larger and larger, but if the income tax is excessively progressive $f' [g(x)]$ is smaller and smaller and the product may also be rather small. In this case there would be not sufficient motivation to work more.

• *Derivative of a quotient*. Applying the chain rule we can now calculate the derivative of a function

$$h(x) = \frac{1}{g(x)} = [g(x)]^{-1}.$$

We have

$$h'(x) = -[g(x)]^{-2} \cdot g'(x) = -\frac{g'(x)}{[g(x)]^2},$$

and therefore

$$\begin{aligned} \mathrm{D}\left[\frac{f(x)}{g(x)}\right] &= \mathrm{D}\left(f(x) \cdot [g(x)]^{-1}\right) = f'(x)[g(x)]^{-1} - f(x)\frac{g'(x)}{[g(x)]^2} = \\ &= \frac{f'(x) \cdot g(x) - f(x) \cdot g'(x)}{[g(x)]^2}. \end{aligned}$$

5.4.2 The derivative of the inverse function

If a function f realizes a one-to-one correspondence between two intervals (a, b) and (c, d), we can speak about the inverse function of f, which we write as $g = f^{-1}$. We recall that f and g are linked by the two identities

$$g[f(x)] = x \qquad f[g(y)] = y \tag{5.16}$$

where x varies in (a, b) and y varies in (c, d).

We suppose that both f and g have derivatives. If we compute the derivative of the first equation in (5.16) and we use the chain rule, we obtain

$$g'[f(x)] \cdot f'(x) = 1.$$

This implies in particular that $f'(x)$ cannot be zero. We can therefore divide this equation by $f'(x)$ and we obtain the *derivative formula for the inverse function*:

$$\boxed{g'(y) = \frac{1}{f'(x)}} \tag{5.17}$$

where we may replace $y = f(x)$ or $x = g(y)$.

Theorem 4.2. *If $f : (a, b) \to (c, d)$ is invertible and has a derivative $f'(x) \neq 0$ for every x in (a, b), then its inverse function $g = f^{-1} : (c, d) \to (a, b)$ also has a derivative for every y in (c, d) and formula (5.17) holds.*

If we consider $x = g(y)$ using Leibniz's notation, formula (5.17) becomes

$$\frac{\mathrm{d}x}{\mathrm{d}y} = \frac{1}{\dfrac{\mathrm{d}y}{\mathrm{d}x}} \qquad !!!$$

Formula (5.17) has a simple geometric meaning. Recall that the graphs of f and f^{-1} are symmetric with respect to the bisector of the first and third quadrants of the Cartesian plane. Therefore, given a pair of corresponding points on the two graphs, the tangent lines at these points (each tangent to the corresponding graph) form complementary angles with the horizontal axis (*i.e.*, if one angle is α the other is $\pi/2 - \alpha$). It thus follows that the respective slopes are reciprocal[3].

[3] We recall that $\tan\left(\frac{\pi}{2} - \alpha\right) = \frac{1}{\tan\alpha}$.

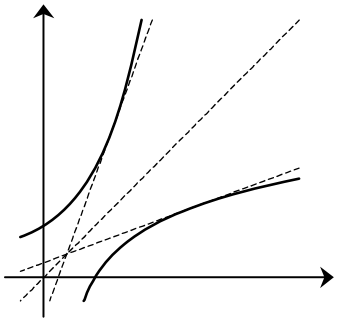


Figure 5.4. Graphs of f and f^{-1} with tangents at two corresponding points

Examples

4.4. We find again a known result. We consider the function $y = f(x) = x^3$, whose derivative $f'(x) = 3x^2$ is nonzero for $x \neq 0$. By the previous theorem we deduce that the inverse function $x = g(y) = \sqrt[3]{y}$ has derivative for $y \neq f(0) = 0$. From (5.17) we have then

$$g'(y) = \frac{1}{f'(x)} = \frac{1}{3x^2} = \frac{1}{3\sqrt[3]{y^2}}.$$

4.5. We calculate the derivative of the function $y = f(x) = \arctan x$. Its inverse function $x = \tan y$ is defined and has a derivative in $(-\pi/2, \pi/2)$, the derivative being $g'(y) = 1 + (\tan y)^2 \neq 0$. The function f therefore has a derivative, which is

$$f'(x) = \frac{1}{g'(y)} = \frac{1}{1 + (\tan y)^2} = \frac{1}{1 + x^2}.$$

5.5 The differential

We go back to the example of the parabola in section 1. The straight line (5.5) provides a good linear approximation to the parabola near x_0: it is, in fact, the best possible linear approximation. Let us try to convince ourselves. Equation (5.4) may be rewritten in the form

$$\Delta f = 2x_0(x - x_0) + (x - x_0)^2. \tag{5.18}$$

If we limit ourselves to considering $|x - x_0|$ to be very small, that is when x is very close to x_0, then $(x - x_0)^2$ is much smaller than $|x - x_0|$. The term $2x_0(x - x_0)$, which is linear in $x - x_0$, then provides a good approximation of Δf. For very small $|x - x_0|$, we can therefore write

$$\Delta f = \underbrace{2x_0(x - x_0)}_{\text{linear term}} + \quad E \tag{5.19}$$

where E stands for "Error" and represents a term which is very small w.r.t. $(x - x_0)$, when $|x - x_0|$ is very small. Now, let us use a computer to investigate the meaning of formula (5.19).

• *A computer experiment.* We take the point with coordinates $(1, 1)$ on the parabola and we draw the tangent line to the graph at this point, which has equation

$$y = 1 + 2\,(x - 1) = 2x - 1.$$

We analyze the same graph at points which are nearer and nearer to $(1, 1)$, recursively zooming the picture. The effect is shown in Figure 5 and constitutes a real experimental proof of formula (5.19). Notice how the parabola is approximated very well by the line (5.5) near $x_0 = 1$. The closer we are to x_0, the better the approximation.

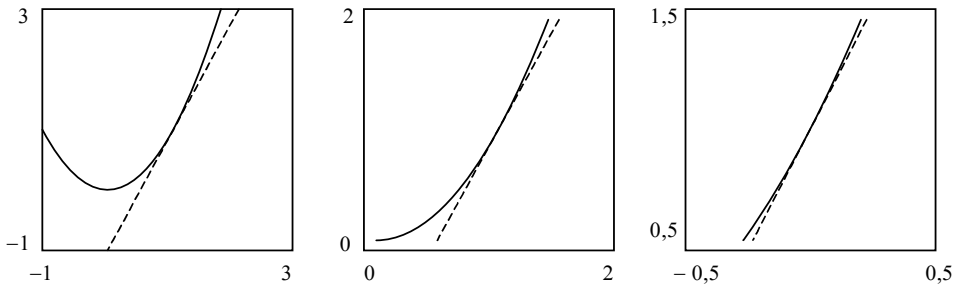


Figure 5.5. Zoom effect

Formula (5.19) expresses these considerations more precisely, and shows in what sense the tangent line approximates the parabola. Can we obtain the same conclusion in the case of a generic function $y = f\,(x)$? The answer is positive if we limit ourselves to "smooth" functions.

We invite the student to perform other computer experiments, using for example the function $y = e^x$ near the point $(1, e)$ or the point $(0, 1)$. The obtained effect is very similar to the one we observed in the case of the parabola. If we zoom more and more, we obtain graphs which are better and better approximated by the tangent line.

It should be noticed that, on the contrary, if we repeat the experiment with the function $f\,(x) = |x|$ and we zoom as much as we want with center in $(0, 0)$, we always obtain the same graph!

We can now draw some conclusions. In the formula (5.19) lies the core concept of *differentiability*: the increment

$$\Delta f = f(x) - f(x_0)$$

of a function $y = f(x)$ can be written as the sum of

— a term which is proportional to $(x - x_0)$, with proportionality coefficient m depending in general on x_0, and

— a term E which is much smaller w.r.t. $(x - x_0)$ when x is near enough to x_0.

In formulae: if $|x - x_0|$ is sufficiently small,

$$f(x) - f(x_0) = m(x_0)(x - x_0) + \ E \tag{5.20}$$

The formula (5.20) is the synthesis of our computer experiments and shows the fact that, if $|x - x_0|$ is sufficiently small, the graph of f "blends in" with the line

$$y = f(x_0) + m\ (x - x_0). \tag{5.21}$$

We can identify the error E in (5.20) by expressing in a rigorous way the fact that, if x is sufficiently near to x_0, the error E in (5.20) is much smaller than $x - x_0$. The translation into formulae requires the notion of limit and it is simply:

$$\lim_{x \to x_0} \frac{E}{x - x_0} = 0. \tag{5.22}$$

The formula (5.22) shows that the error we commit by substituting the tangent line to the original graph is an infinitesimal of higher order with respect to $x - x_0$, that is $E = o(x - x_0)$. Formula (5.20) therefore becomes

$$f(x) - f(x_0) = m(x - x_0) + o\,(x - x_0)\,, \qquad \text{for } x \to x_0. \tag{5.23}$$

Definition 5.1. *Let f be defined in the interval (a, b) and $x_0 \in (a, b)$; f is said to be* ***differentiable*** *at x_0 if there exists a real number $m = m\,(x_0)$ such that* (5.23) *holds. The linear term $m(x - x_0)$ is called the* (*first order*) ***differential*** *of the function at x_0, relative to the increment $x - x_0$.*

Even if the differential does not coincide with the increment Δf, it captures a large part of the increment[4] provided that x is sufficiently near to x_0.

Remark. By setting $h = x - x_0$, formula (5.23) can be rewritten in the form

$$f(x_0 + h) - f(x_0) = m\,(x_0)\,h + o\,(h)\,, \qquad \text{for } h \to 0.$$

We can say that a function f is differentiable if the increment Δf (regarded as a function of h) is well approximated by the linear function

$$h \longmapsto m\,(x_0)\,h$$

(which is the differential of f evaluated in h).

Derivative and Differential

Let us identify the straight line (5.21). In the case of the parabola, it coincides with the tangent line and therefore $m = f'\,(x_0)$. This always happens for differentiable functions and therefore a differentiable function also has a derivative. Conversely, if a function has a derivative, then the formula (5.23) holds with $m = f'\,(x_0)$.

[4] Of course, it may happen that $m(x_0) = 0$, and in this case the differential is zero and "the large part" of the value of Δy is zero! This means that, for x near to x_0, the function is approximated by the constant value $f(x_0)$ up to infinitesimals of order higher than $(x - x_0)$.

Theorem 5.1. *f is differentiable at x_0 if and only if f has a derivative at x_0. We can then write, for $x \to x_0$,*

$$f(x) - f(x_0) = f'(x_0)(x - x_0) + o(x - x_0) \tag{5.24}$$

Proof. If f is differentiable, formula (5.23) holds; we divide it by $(x - x_0)$, obtaining

$$\frac{f(x) - f(x_0)}{x - x_0} = m \ + o(1)$$

and we pass to the limit for $x \to x_0$. Since $o(1) \to 0$, we deduce that

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = m$$

and therefore f has a derivative, precisely $f'(x_0) = m$. Conversely, if f has a derivative, the relation

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0)$$

is equivalent to

$$\frac{f(x) - f(x_0)}{x - x_0} = f'(x_0) + o(1) \qquad \text{for } x \to x_0.$$

Multiplying by $x - x_0$, and since[5] $(x - x_0)\, o(1) = o(x - x_0)$, we find that

$$f(x) - f(x_0) = f'(x_0)(x - x_0) + o(x - x_0) \qquad \text{for } x \to x_0,$$

which is the differentiability condition with $m = f'(x_0)$. □

Example 5.1. The functions $f_1(x) = \sin x$, $f_2(x) = \ln(1 + x)$ and $f_3(x) = e^x - 1$, as the examples 6.5 and 6.6 of chapter 3 show, are differentiable at $x_0 = 0$. Indeed, for $x \to 0$ we can write

$$\begin{aligned} \sin x &= x + o(x) \\ \ln(1 + x) &= x + o(x) \\ e^x - 1 &= x + o(x). \end{aligned}$$

The differentials obey the rules of calculation which we can immediately deduce from the analogous ones holding for derivatives. In particular, for the sum, the product and the quotient, if f and g are differentiable, we have

$$\mathrm{d}(f + g) = \mathrm{d}f + \mathrm{d}g, \quad \mathrm{d}(f \cdot g) = g\mathrm{d}f + f\mathrm{d}g, \quad \mathrm{d}\left(\frac{f}{g}\right) = \frac{g\mathrm{d}f - f\mathrm{d}g}{g^2}.$$

[5] By definition $\dfrac{o(x - x_0)}{x - x_0} \to 0$ for $x \to x_0$.

• *The differential.* In general, when we work with differentials, the increment $|x - x_0|$ is very, very small: we can therefore allow ourselves to call it "*infinitesimal*" and denote it by the symbol $\mathrm{d}x$ (in contrast to Δx, which denotes *any* increment).

The differential, if non-zero, is the most consistent part of the increment $\Delta f = f(x) - f(x_0)$ and is usually denoted by the symbol $\mathrm{d}f(x_0)$. Following what we have previously written, it is expressed by the formula

$$\boxed{\mathrm{d}f(x_0) = f'(x_0)\mathrm{d}x}. \tag{5.25}$$

This justifies the notation for the derivative (introduced by Leibniz):

$$f'(x_0) = \frac{\mathrm{d}f(x_0)}{\mathrm{d}x}.$$

According to this formula, the derivative $f'(x_0)$, or the slope of the graph of f at x_0, appears as the rate of variation of f relative to an *infinitesimal* variation of x. This way of thinking proves to be effective in the applications of calculus. At the same time, the formula (5.25) shows that the differential of f at x_0 is the *linear function*

$$\mathrm{d}x \mapsto f'(x_0)\mathrm{d}x.$$

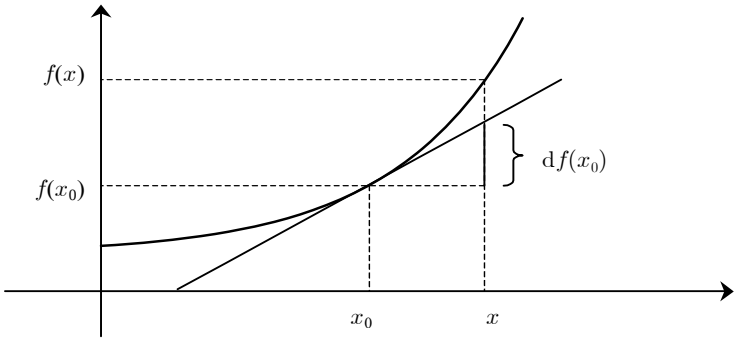


Figure 5.6. Geometric meaning of the differential

Invariance of form of the differential

The differential of the composite function $z = f[g(x)]$ is

$$\mathrm{d}z = f'[g(x)]g'(x)\mathrm{d}x. \tag{5.26}$$

It may be interesting to observe that, since $y = g(x)$ and therefore $\mathrm{d}y = g'(x)\mathrm{d}x$, the formula (5.26) can be rewritten in the equivalent form

$$\mathrm{d}z = f'(y)\mathrm{d}y, \tag{5.27}$$

which coincides with the differential of $z = f(y)$ *when* y *is considered to be an independent variable*.

In conclusion, the differential of a function $z = f(y)$ is represented by formula (5.27) *both when* y *is an independent variable and when* y *depends in its turn on another variable*. This property, known as *the invariance of form of the differential*, allows us to differentiate an equation containing different variables without needing to specify *a priori* which variables are to be considered as independent.

Example 5.2. We consider again the example of the factory given at the beginning of section 4 (cfr. page 131) and we consider a variation $\mathrm{d}q$ of the produced quantity. The variation of the cost function is well approximated by the differential $C'(q)\mathrm{d}q$. If the variation of q is determined by the variation of the price x, so that $\mathrm{d}q = q'(x)\mathrm{d}x$, it is sufficient to substitute this expression in the differential of the cost to obtain the differential of the cost relative to the price variation:

$$C'(q)\,\mathrm{d}q = C'(q)\,\underbrace{q'(x)\,\mathrm{d}x}_{\mathrm{d}q}.$$

5.6 Elasticity and semi-elasticity

5.6.1 Elasticity

In the analysis of many models in Economics where real-life functions come into play, for example the study of the demanded amount of goods as a function of the price, it is often more meaningful to consider, instead of the absolute increments of the variables x and $y = f(x)$ (that is h and $f(x_0 + h) - f(x_0)$), the corresponding relative increments (or percentage increment) :

$$\frac{h}{x_0} \quad \text{and} \quad \frac{f(x_0 + h) - f(x_0)}{f(x_0)}.$$

There are two main reasons. The relative increments show in a clearer way the importance of the considered variations w.r.t. the initial value of the variable. For example, if x denotes a price, an increment h of 100 has a completely different importance if the initial price is $x_0 = 500$ or if it is $x_0 = 10000$: the percentage increment is respectively 20% or 1%, having a completely different relevance.

A second reason lies in the fact that relative increments are pure numbers, that is they do not depend on the chosen unit of measure. Let us suppose again that X is a price. An increment of $h = 100$ has a clearly different importance if the price is expressed in Euro, or in Yen, or in US Dollars. On the contrary, the statement that the price has increased by 100% has the same meaning independently of the currency used.

Having considered these preliminary remarks, we define as *arc elasticity* of a positive function f at a point $x_0 > 0$ relative to an increment h of the variable x, the

ratio of the relative increments

$$\frac{\dfrac{f(x_0+h)-f(x_0)}{f(x_0)}}{\dfrac{h}{x_0}} = \frac{f(x_0+h)-f(x_0)}{h}\cdot\frac{x_0}{f(x_0)}. \tag{5.28}$$

Therefore, the arc elasticity is the difference quotient of f, multiplied by the quotient $x_0/f(x_0)$. If $h \to 0$, we obtain the *point elasticity*.

Definition 6.1. *We call* ***point elasticity*** *(or simply* ***elasticity****) of f at x_0, denoted by the symbol $E_f(x_0)$, the product*

$$f'(x_0)\cdot\frac{x_0}{f(x_0)}.$$

This product can be rewritten in a more relevant form, as the ratio between the marginal value of f and the average unitary value. That is, we have

$$\boxed{E_f(x_0) = \frac{f'(x_0)}{\dfrac{f(x_0)}{x_0}}}$$

A function f is said to be *elastic* at x_0 if $|E_f(x_0)| > 1$, and *inelastic* if $|E_f(x_0)| < 1$. If $|E_f(x_0)| = 1$, some authors say that f is *anelastic*. Given a relative variation of x_0, when f is inelastic this variation causes a relative variation of $f(x_0)$ which is less than that of x_0, whereas when f is elastic the relative variation of $f(x_0)$ is larger than that given for x_0.

We also have

$$E_f(x_0) = \frac{\mathrm{D}\ln f(x_0)}{\mathrm{D}\ln x_0}.$$

For differentiable positive functions the *elasticity function* is therefore well defined:

$$E_f(x) = f'(x)\frac{x}{f(x)}.$$

Example 6.1. Let $f(x) = 1/x^k$, with $k > 0$. We have:

$$E_f(x) = -k$$

Thus if $k > 1$ the function is everywhere elastic, if $k < 1$ everywhere inelastic, if $k = 1$ everywhere anelastic.

We can also determine whether f is elastic or inelastic at a given point by means of geometric considerations about its graph. We consider, for simplicity, the case of an increasing function: its derivative $f'(x_0)$ represents the slope of the tangent line t to the graph of f at $(x_0, f(x_0))$, while $f(x_0)/x_0$ is the slope of the line r passing through the origin and the point $(x_0, f(x_0))$. Therefore, the fact that f is elastic at x_0 means that the slope of t is greater than the slope of r, while the fact that f is inelastic means that the slope of t is less than the slope of r.

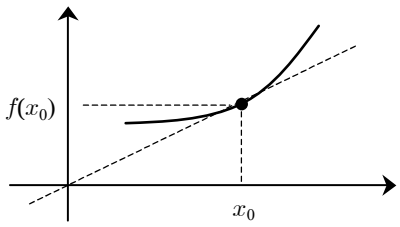


Figure 5.7. A point at which the elasticity is equal to one

Proposition 6.1. *At every point where the elasticity of f is equal to one, the tangent to the graph passes through the origin.*

The rules of calculation for elasticity can be deduced from the rules we have seen before. We invite the reader to verify that *the elasticity of a product is the sum of the elasticities*.

Example 6.2. We calculate the elasticity of a power function and of an exponential function. Let

$$f(x) = Ax^\alpha \qquad A > 0,\ \alpha \text{ real},\ x > 0.$$

We have

$$E_f(x) = A\alpha x^{\alpha-1}\frac{x}{Ax^\alpha} = \alpha.$$

The power functions have constant elasticity. Let now

$$f(x) = Ae^{\alpha x} \qquad A > 0,\ \alpha \text{ and } x \text{ real}.$$

We have

$$E_f(x) = A\alpha e^{\alpha x}\frac{x}{Ae^{\alpha x}} = \alpha x.$$

The exponential functions have linear elasticity.

• *Elasticity of demand*. Let $q = q(p)$ be the demand function of some goods depending on the price p. For the *standard* goods we assume that q' is negative (i.e., as we shall see later, that the demand is a decreasing function of the price). In this case we often prefer to define elasticity as

$$E_q(p) = -q'(p)\frac{p}{q(p)},$$

that is we consider the absolute value of the usual elasticity, which we defined previously. Thus, for example, an elasticity of 2 means that an increment (reduction) in price of 10% causes a market reaction of a reduction (increment) in demand of 20%, while an elasticity of 1/2 means that the same increment in price (reduction) causes a reduction (increment) in demand of 5%.

5.6.2 Logarithmic derivative or semi-elasticity

Another aspect of the elasticity of a function f can be noticed from the formula for the derivative of a composite function. We start by defining the *logarithmic derivative* of a *positive* function f by the formula

$$\mathrm{D}\ln f(x) = \frac{f'(x)}{f(x)}.$$

Its meaning should be evident: it expresses the relative rate of increment of f w.r.t. x. Indeed, it is also equal to the limit of the ratio between the percentage increment of f and the increment of x, when the increment of x tends to 0:

$$\lim_{h\to 0} \frac{\dfrac{f(x+h)-f(x)}{f(x)}}{h}.$$

Such a limit defines the *point semi-elasticity* of f. An interesting application can be found in the chapter on Financial Calculus (⇛ **Chapter 11**): it is called the *force of interest* of a financial law.

Whenever we are interested in visualizing the relative increments of a given function, we generally use the so-called *semi-logarithmic* scale: as usual, the horizontal axis represents the values of x, while the vertical one represents the values of $Y = \ln f(x)$[6]. In such a graph the slope of the tangent is the semi-elasticity, that is the rate which we are considering now.

We use the semi-logarithmic scale also when f grows so fast that it requires a strong compression of the vertical axis. For example, in semi-logarithmic scale the graph of the power function $f(x) = Ax^\alpha$ becomes the logarithmic curve with equation $Y = \ln A + \alpha \ln x$, while the graph of the exponential function $f(x) = Ae^{\alpha x}$ becomes the line of equation $Y = \alpha x + \ln A$.

For positive functions, defined for $x > 0$, we can also use a representation in *logarithmic scale*: on the horizontal axis we have the values of $X = \ln x$, while on the vertical one we place the values of $Y = \ln f(x)$.

In this scale, the graph of the power function we considered before becomes a line. This can be observed by applying a logarithm function to both sides of the equation $y = Ax^\alpha$ (with the original variables), thus obtaining $\ln y = \ln A + \alpha \ln x$, and then substituting $Y = \ln y$, $X = \ln x$. The resulting equation is affine and linear:

$$Y = \alpha X + \ln A.$$

Using the same computations we see that the graph of the exponential function $f(x) = Ae^{\alpha x}$ becomes the graph of a translated exponential function: $Y = \ln A + \alpha e^X$.

Proposition 6.2. *The slope of a graph in logarithmic scale is the elasticity.*

[6] We can also find graphs in semi-logarithmic scale which use logarithms with base 10.

Proof. With the chain formula, using $y = \ln x$ as intermediate variable, we have

$$\frac{\mathrm{d}\ln f(x)}{\mathrm{d}x} = \frac{\mathrm{d}\ln f(x)}{\mathrm{d}\ln x}\frac{\mathrm{d}\ln x}{\mathrm{d}x}. \tag{5.29}$$

Since

$$\frac{\mathrm{d}\ln f(x)}{\mathrm{d}x} = \frac{f'(x)}{f(x)}, \qquad \frac{\mathrm{d}\ln x}{\mathrm{d}x} = \frac{1}{x},$$

from (5.29) we deduce that

$$\frac{f'(x)}{f(x)} = \frac{\mathrm{d}\ln f(x)}{\mathrm{d}\ln x}\frac{1}{x}$$

and therefore the slope of the graph of f in logarithmic scale satisfies the asserted relationship:

$$\frac{\mathrm{d}\ln f(x)}{\mathrm{d}\ln x} = x\frac{f'(x)}{f(x)} = E_f(x). \ \square$$

• (⇛ **Chapter 11**) *Duration.* We consider a share which pays the amount a after t years. If the compound interest rate r holds, its value today is

$$P(r) = \frac{a}{(1+r)^t},$$

since if we buy it we will get a at time t. Its market price cannot be different from this value[7]. This formula tells us that the price of the share is equal to its *discounted value*.

If a more complex bond pays a_1 after t_1 years, a_2 after t_2 years,... , a_n after t_n years, and if the compound interest rate r is the same for each maturity $t_1, t_2, \ldots, t_n$, its price is the sum of the discounted values of all entries that the bond provides:

$$P(r) = \sum_{s=1}^{n} \frac{a_s}{(1+r)^{t_s}}.$$

The price of the share depends on the market interest rate r. Can we determine the relative variation $\Delta P/P$ of the share price which originates from a variation Δr of the market interest rate?

Since rates are "small" numbers, their variations are also small and therefore the differential $\mathrm{d}[\ln P(r)] = \mathrm{d}P(r)/P(r)$ provides a good estimate of the relative variation of the price w.r.t. the variation of the interest rate. We have

$$P'(r) = \sum_{s=1}^{n} -\frac{t_s a_s}{(1+r)^{t_s+1}}$$

and therefore

$$\frac{\mathrm{d}P(r)}{P(r)} = -\frac{D}{1+r}\mathrm{d}r \tag{5.30}$$

[7] Otherwise, an arbitrage would be possible, that is the market would react with speculations, which would bring the situation back to normal in a very short time.

where the number

$$D = \frac{\sum_{s=1}^{n} t_s a_s (1+r)^{-t_s}}{\sum_{s=1}^{n} a_s (1+r)^{-t_s}}$$

is called the *duration* of the cash flow which is generated from the considered share. The ratio

$$D^* = \frac{D}{1+r}$$

is sometimes called the *modified duration* and measures the sensitivity of price w.r.t. the interest rate, because the formula (5.30) tells us that

$$\boxed{\text{percentage variation of price} \simeq \ -D^* \times \text{variation of rate.}}$$

Let us see how things work with a concrete example. A share pays the amounts which are shown in the following table:

Maturity	Amount
1	10
2	110

If the interest rate r is 10%, the current price of this share is

$$P(10\%) = \frac{10}{1.1} + \frac{110}{1.21} = 100.$$

We study the effects of an increment $\Delta r = 1\%$. The price variation we actually get is

$$P(11\%) - P(10\%) = \frac{10}{1.11} + \frac{110}{1.11^2} - 100 = -1.7125,$$

therefore the percentage variation is -1.7125%. Let us see how this rate can be well approximated via the modified duration. The *duration* of the cash flow is

$$\frac{1 \times \frac{10}{1.1} + 2 \times \frac{110}{1.21}}{100} = 1.9091$$

and therefore (using the formula described above)

$$\frac{P(11\%) - P(10\%)}{P(10\%)} \simeq -\frac{1.9091}{1.1} \times 0.01 = -1.7355\%$$

which, as we see, differs very little from the true variation -1.7125%.

5.7 Optimization and stationary points

Differential calculus is the main tool for the solution of optimization problems. We introduce here the most elementary theoretical aspects by illustrating a problem of a "geometric-economic" nature.

• *The correct proportions of a beer can.* The managers of the beer factory Schiumador realize that their cans cost too much. They must maintain the cylindrical shape, the material, and the volume (33cl). So they have to act on the proportions of the can, trying to minimize, in particular, the total surface of the can: this is in fact directly linked to the amount of material needed to produce the can. The problem is: how to choose the height and the base radius in such a way that the total surface is minimal, while maintaining a constant volume of 33cl?

Let V be the volume of the can, h its height, r its base radius. We recall the formulae

$$V = \pi h r^2, \qquad h = \frac{V}{\pi r^2}.$$

Therefore, the total surface area is

$$A(r) = \underbrace{2\pi r^2}_{\text{base surface area}} + \underbrace{2\pi h r}_{\text{side surface area}} = 2\pi r^2 + \frac{2V}{r}.$$

We observe that the two terms of the sum are "in competition" with each other: small values of r produce small base areas, but a large side area. The contrary holds for large values of r. We are therefore inclined to believe in the existence of a value r^* which balances the two terms and minimizes the sum.

How can we find such a value r^*? We plot the graph of the function $A(r)$: we expect a behaviour of the type shown in Figure 8.

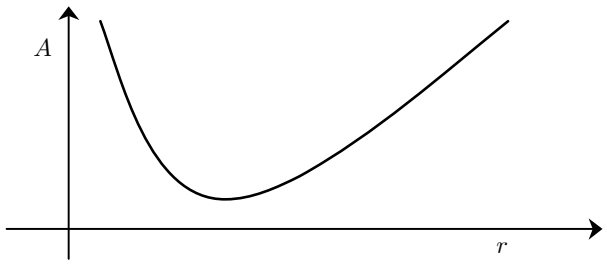


Figure 5.8. Surface area as a function of the base radius

For small values of r (phial-like cans) $A(r)$ takes large values (because of the term $2V/r$), while for large values of r (film-reel-like cans) $A(r)$ again takes large values (this time because of the term $2\pi r^2$). The function A is therefore decreasing up to a certain value r^* and increasing from r^* onwards: r^* will be a point of minimum for $A(r)$.

What characterizes, from a geometric point of view, the point of minimum r^*?

The tangent to the graph of A at such a point is horizontal.

In other terms, the slope of the graph of $A(r)$ at r^* is 0! We therefore have a method for determining r^*: we compute the derivative of the function A and set it to be 0. We obtain the equation

$$A'(r) = 4\pi r - \frac{2V}{r^2}.$$

By setting $A'(r) = 0$, we obtain

$$4\pi r - \frac{2V}{r^2} = 0,$$

and therefore

$$4\pi r^3 - 2V = 0$$

and finally we obtain the required value:

$$r^* = \sqrt[3]{\frac{V}{2\pi}}.$$

Are the cans on the market "optimized"? The common 33 cl cans have height equal to 11.5 cm and base diameter equal to 6.5 cm, thus the base radius is $r = 3.25$. The volume is therefore

$$V = 3.25^2 \times \pi \times 11.5 \simeq 382 \text{ cm}^3.$$

The optimized radius is approximately

$$r^* = \sqrt[3]{\frac{382}{2\pi}} \simeq 3.93$$

The optimized can is therefore a little less comfortable to hold, but it has lower costs.

We invite the reader to check that the "optimal can" is the one with base diameter equal to the height ($2r = h$).

The simple geometric considerations which allowed us to solve the problem can be generalized and reformulated in a rigorous way in the following fundamental theorem: *Fermat's Theorem*[8].

Theorem 7.1. *Let f be a function defined in (a, b). If f has a point of (local) maximum or minimum at a point x_0, where f is differentiable, then*

$$\boxed{f'(x_0) = 0.} \tag{5.31}$$

Formula (5.31) is a so-called *first order condition*[9], and it is a necessary condition in order that x_0 be a point of local extremum. The points at which the derivative of

[8] Pierre de Fermat (1601-1665).

[9] Because it is a condition on the first derivative, the derivative of the function. In the international economic literature such a condition is often called *FOC* (*First Order Condition*).

a function is null (the x-coordinates of the points where the graph has a horizontal tangent) are special ones, and are called *stationary points*.

Proof. Let us consider the case in which f has a *local minimum* at x_0 (the other one is similar). This means that, moving a little to the right or to the left of x_0, the values of f are in any case greater or equal to $f(x_0)$. In formulae we can write that, for $|h|$ small enough, $f(x_0 + h) - f(x_0) \geqslant 0$. We then have:

$$\frac{f(x_0 + h) - f(x_0)}{h} \geqslant 0 \qquad \text{for } h > 0 \tag{5.32}$$

and

$$\frac{f(x_0 + h) - f(x_0)}{h} \leqslant 0 \qquad \text{for } h < 0. \tag{5.33}$$

We now pass to the limit for $h \to 0$. Remember that f has a derivative at x_0 and that the limit preserves the inequalities (not the strict ones!): from (5.32) we obtain $f'(x_0) \geqslant 0$, while from (5.33) we obtain $f'(x_0) \leqslant 0$. We therefore deduce that $f'(x_0) = 0$. □

Remarks

1. The fact that a point x_0 is a stationary point for a function$f$ does not allow us to think that it is automatically a maximum or minimum point. Indeed, the function $f(x) = x^3$ has a stationary point $x = 0$ since $f'(x) = 3x^2$ is equal to zero for $x = 0$, but this point is neither a maximum nor a minimum point. In the example of the beer can, the condition $f'(x) = 0$ could be regarded as sufficient, since we already knew that there was a minimum point by looking at the graph of f: we simply observed that the minimum point had to be stationary, and this was enough to find it ($f'(x) = 0$ had a unique solution). Generally speaking, this condition is only used to select *the points which are candidates to be extremum points*, and then other information is needed to recognize their nature.

2. In the statement of Fermat's theorem, the interval is without extrema, that is open. This is not fortuitous: if we choose an interval including one or both extrema, the conclusion is different. For example let x_0 be a maximum point for f, a differentiable function in the *closed* interval $[a, b]$. If x_0 is the extremum a the function may have a negative right derivative, and if x_0 is the extremum b the function may have a positive left derivative; Figure 9 shows these two situations.

In any case, if f is a differentiable function in $[a, b]$ and f has a maximum at a point x_0 which is not in (a, b) (thus $x_0 = a$ or $x_0 = b$), by looking at Figure 9 we can still say that necessarily

$$f'(x_0)(x - x_0) \leq 0.$$

Analogously, if x_0 is a minimum point for f, then necessarily

$$f'(x_0)(x - x_0) \geq 0.$$

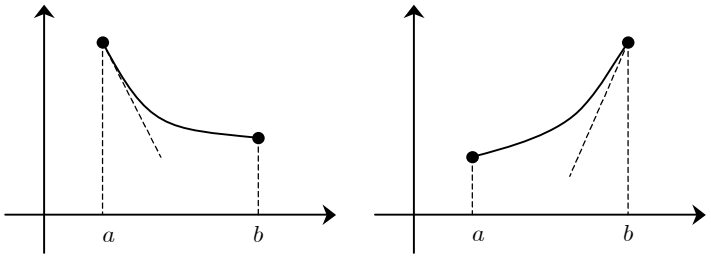


Figure 5.9. Non-stationary maximum points

3. We want to stress once more that a point can be an extremum point for f even if f is not differentiable there. For example, $f(x) = |x|$ is not differentiable at $x = 0$, but has a minimum point there. In an extremum point, f may even be discontinuous!

4. In the case of *optimization on a closed and bounded interval* $[a, b]$, the condition expressed by Fermat's theorem is often sufficient to determine the global maximum and minimum, which exist because of Weierstrass's theorem. Let us explain this better with an example. We want to compute the global maximum and minimum of the function

$$f(x) = x\sqrt{1-x^2}$$

in the interval $[0, 1]$. The function is continuous there, hence both extrema exist. Since $f(0) = f(1) = 0$, while for $x \in (0, 1)$ the function is positive, the minimum of f is 0 (thus $x = 0$ and $x = 1$ are minimum points). The points of maximum will belong to the open interval $(0, 1)$: f is differentiable there and by Fermat's theorem f' must be 0. We have:

$$f'(x) = \sqrt{1-x^2} + \frac{x}{2\sqrt{1-x^2}}(-2x) = \frac{1-2x^2}{\sqrt{1-x^2}}$$

and therefore $f'(x) = 0$ in $(0, 1)$ only for $x = 1/\sqrt{2}$. This unique stationary point must therefore necessarily be the maximum point, and the maximum of f is $f\left(1/\sqrt{2}\right) = 1/2$.

We conclude with two examples from Economics.

• *A problem of efficiency.* Let, as usual, $C = C(q)$ be the production cost of a factory for producing a quantity q of goods. The efficiency of the production process can be measured by the ratio $C(q)/q$, which is the average cost for a unit of goods. The maximum efficiency is reached when the average cost is minimum.

A possible curve for the average cost, with only one minimum point, could be the one in Figure 10.

If the cost function is *smooth* and we denote by q_0 the point of maximum efficiency, then, because of Fermat's theorem, the derivative of the average cost must be zero

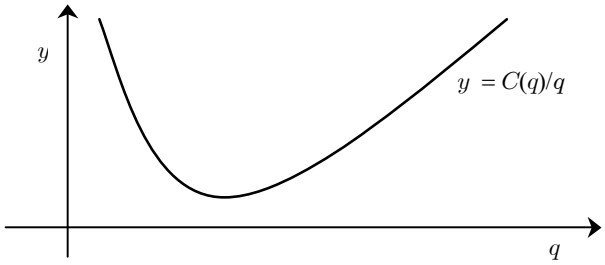


Figure 5.10. A curve of average cost

at q_0. By the formula for the derivative of a quotient we must then have

$$\frac{C'(q_0)q_0 - C(q_0)}{q_0^2} = 0.$$

Setting the numerator to be zero, this condition can be rewritten in the two following forms, which are the most meaningful ones:

$$C'(q_0) = \frac{C(q_0)}{q_0}, \qquad \frac{q_0 C'(q_0)}{C(q_0)} = 1.$$

For the production amount of maximum efficiency we must therefore have

$$\boxed{\text{marginal cost} = \text{average cost}}$$

or equivalently

$$\text{elasticity of } C \text{ at } q_0 = E_C(q_0) = 1.$$

• *Maximum revenue.* Let $q = q(p)$ be the demand function of one sort of goods as a function of the price p. Then the function $r(p) = pq(p)$ expresses the revenue as a function of the price. Let us suppose that there exists a price p_0, which maximizes the revenue. If the revenue function is differentiable and p_0 is neither the minimum nor the maximum admissible price, then Fermat's theorem tells us that $r'(p_0) = 0$. But we have

$$r'(p) = q(p) + pq'(p),$$

hence the stationary condition of the revenue can be written as

$$q(p_0) + p_0 q'(p_0) = 0$$

and thus also as

$$\text{elasticity of } q \text{ at } p_0 = -p_0 \frac{q'(p_0)}{q(p_0)} = E_q(p_0) = 1.$$

5.8 Lagrange's mean value theorem

Let us suppose that a car travels with regularity for 3 hours and covers 180km, thus travelling at an average speed of 60 km/h. Surely, at some instant the speed was exactly 60 km/h. We mean that at some time the instantaneous velocity was equal to the average velocity.

To express the same concept in geometric terms, we consider the graph of a *smooth* function f which joins two points of coordinates $(a, f(a))$ and $(b, f(b))$, as shown in figure 11.

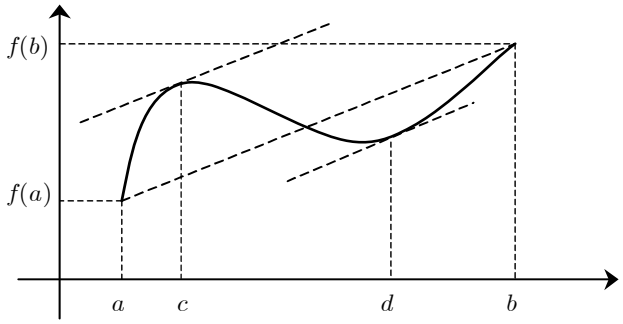


Figure 5.11. Geometric meaning of Lagrange's mean value theorem

As we see, there exists at least one point in the graph of f (in Figure 11 there are two) where the tangent line is parallel to the line joining the extrema. *Lagrange's mean value theorem* (not to be confused with the other "Mean value theorem" which we shall see in Integral Calculus) translates this parallelism condition and has many important consequences.

Theorem 8.1 (Lagrange's mean value theorem). *Let f be a differentiable function in an interval (a, b), which is continuous also in a and b; then there exists at least one point c in (a, b) such that*

$$f'(c) = \frac{f(b) - f(a)}{b - a}. \tag{5.34}$$

Proof. The line joining the two points $(a, f(a))$ and $(b, f(b))$ has equation

$$y = f(a) + \frac{f(b) - f(a)}{b - a}(x - a),$$

which is an affine linear function. We consider the difference between f and such function, that is

$$g(x) = f(x) - f(a) - \frac{f(b) - f(a)}{b - a}(x - a).$$

We note that g is zero at the extrema a, b, differentiable in (a, b) and

$$g'(x) = f'(x) - \frac{f(b) - f(a)}{b - a}.$$

The theorem's assertion corresponds to the statement that there exists a point c in (a, b) where the derivative of g is zero. Now, g is continuous in the interval, extrema included; Weierstrass's theorem implies that g has a maximum M and a minimum m. These are in correspondence with at least two points, say c_1 and c_2, such that $g(c_1) = M$ and $g(c_2) = m$.

If c_1 and c_2 are both extrema of the interval, since g is zero at those points we have $M = m = 0$: in this case g is everywhere zero and the graph of f is the line under consideration, with constant slope as in (5.34) for every $x \in (a, b)$:

$$f'(x) = \frac{f(b) - f(a)}{b - a}.$$

If not, then at least one of the points c_1 and c_2 lies in (a, b). Let us call this point c. Since c is an extremum point, the derivative of g is zero at c by Fermat's theorem. □

The particular case $f(a) = f(b)$ is sometimes known as *Rolle's theorem*[10]:

Corollary 8.2. *Let f be differentiable in (a, b) and continuous also in a and b, with $f(a) = f(b)$. Then there exists at least one point $c \in (a, b)$ such that*

$$f'(c) = 0.$$

Examples

8.1. We apply Lagrange's mean value theorem to the parabola $f(x) = x^2$ in a generic interval $[a, b]$. We have $f(a) = a^2$, $f(b) = b^2$ and $f'(x) = 2x$. By theorem 8.1 there exists a point c such that

$$2c = \frac{b^2 - a^2}{b - a} = b + a.$$

In this case c is the *arithmetic mean* $(a + b)/2$ of the extrema.

8.2. We consider the hyperbola $f(x) = 1/x$ in an interval $[a, b]$ with $a > 0$. We have $f(a) = 1/a, f(b) = 1/b$ and $f'(x) = -1/x^2$. By Theorem 8.1, there exists a number c such that

$$-\frac{1}{c^2} = \frac{\dfrac{1}{b} - \dfrac{1}{a}}{b - a} = -\frac{1}{ab}.$$

In this case c is the *geometric mean* $\sqrt{ab}$ of the extrema.

5.9 Monotonicity test

We have already commented that elementary considerations are not always sufficient for recognizing the nature of a stationary point, that is if this point is a minimum or

[10] Michel Rolle (1652-1719).

maximum point. In general, we can analyze the behaviour of the function near a stationary point. To this purpose, information about the monotonicity of the functions is crucial: if, for example, the function is increasing before the point, and decreasing afterwards, it is clear that the stationary point is a point of local maximum.

Now, the monotonicity of a function is linked to the sign of the difference quotient relative to two points of its domain.

Let f be defined in (a, b) and consider the difference quotient

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1}, \tag{5.35}$$

for any possible pair of points x_1, x_2 in (a, b) with $x_1 \neq x_2$.

To state that f is *monotonic increasing* is equivalent to stating that the ratio (5.35) is always positive (or zero, if f is not strictly increasing), since the numerator and denominator have the same sign. Analogously, to state that the ratio is always negative (or zero) is equivalent to stating that f is *monotonic decreasing*, since this time the numerator and denominator have opposite signs.

If the function is differentiable, passing to the limit in (5.35) for $x_2 \to x_1$ we obtain $f'(x_1)$. We recall that the limit operation maintains the signs $\leq$ and $\geq$: we immediately obtain the following proposition.

Proposition 9.1. *Let f be differentiable in (a, b). If f is increasing, for every $x \in (a, b)$ we have $f'(x) \geq 0$. If f is decreasing, for every $x \in (a, b)$ we have $f'(x) \leq 0$.*

A comment: one is tempted to remove the equality sign in the case where f is strictly increasing or decreasing. This is not correct, as the function $f(x) = x^3$ shows: it is strictly increasing in $\mathbb{R}$, with derivative $f'(x) = 3x^2$, which is not everywhere positive since $f'(0) = 0$.

The previous proposition is an immediate consequence of the definition of derivative and of the property on the permanence of sign under the limit operation. The converse statement, which is frequently applied in practice, is a consequence of Lagrange's mean value theorem.

Theorem 9.2 (Monotonicity test). *Let $f : \mathbb{R} \to \mathbb{R}$ be differentiable in the interval (a, b). If for every $x \in (a, b)$ we have $f'(x) \geq 0$ ($f'(x) > 0$), then f is increasing (strictly increasing) in the interval (a, b).*

Analogously, if for every $x \in (a, b)$ we have $f'(x) \leq 0$ ($f'(x) < 0$), then f is decreasing (strictly decreasing) in the interval (a, b).

The proof is simple and instructive. We limit ourselves to the case

$$f'(x) \geq 0 \quad \Rightarrow \quad f \text{ increasing.}$$

The other cases are perfectly analogous.

Proof. Let $f'(x) \geq 0$ in (a, b) and consider any pair of points $x_1, x_2 \in (a, b)$ with $x_1 < x_2$. In the interval $[x_1, x_2]$, the assumptions of Lagrange's theorem are satisfied and therefore there exists a point $c \in (x_1, x_2)$ such that

$$f'(c) = \frac{f(x_2) - f(x_1)}{x_2 - x_1}.$$

Since $f'(c) \geq 0$, we have that

$$\frac{f(x_2) - f(x_1)}{x_2 - x_1} \geq 0,$$

and therefore, since x_1, x_2 are arbitrary, f is increasing over the whole interval (a, b). □

Example 9.1. We consider the function $f(x) = x + e^{-x}$ in the interval $(0, +\infty)$. This function is differentiable and we have

$$f'(x) = 1 - e^{-x}.$$

For $x > 0$ the second addend is less than 1, thus $f'(x) > 0$ for every $x > 0$. We can conclude that f is strictly increasing in the interval $(0, +\infty)$.

First test for stationary points

From the *monotonicity test* we easily deduce a *sufficient condition* to guarantee that a stationary point is a point of local maximum or minimum.

Theorem 9.3. *Let f be differentiable in the interval (a, b) and let $x_0 \in (a, b)$ be a stationary point, that is a point such that $f'(x_0) = 0$.*

If in a left neighbourhood of x_0 we have $f'(x) \geq 0$ and in a right neighbourhood of x_0 we have $f'(x) \leq 0$, then x_0 is a point of local maximum.

If in a left neighbourhood of x_0 we have $f'(x) \leq 0$ and in a right neighbourhood of x_0 we have $f'(x) \geq 0$, then x_0 is a point of local minimum.

$x < x_0$	$x > x_0$
$f'(x) \geq 0$	$f'(x) \leq 0$
$f \nearrow$	$f \searrow$

$\Rightarrow x_0$ maximum point

$x < x_0$	$x > x_0$
$f'(x) \leq 0$	$f'(x) \geq 0$
$f \searrow$	$f \nearrow$

$\Rightarrow x_0$ minimum point

Example 9.2. We consider the function $f(x) = (x^2 - 8)e^x$, clearly defined and differentiable in $\mathbb{R}$, and we check if it has local maxima and/or minima. First of all, we look for stationary points by setting the derivative of f equal to 0. We have:

$$f'(x) = (x^2 + 2x - 8)e^x.$$

This derivative is zero at $x = -4$ and at $x = 2$, the only stationary points of f. To establish their nature, we analyse the sign of f'. The factor e^x is always positive and therefore it is irrelevant. On the other side, the trinomial $x^2 + 2x - 8$ is positive if $x < -4$ or $x > 2$ and negative if $-4 < x < 2$. The following table summarizes the conclusions we can deduce:

	$x < -4$	$-4 < x < 2$	$x > 2$
f'	$+$	$-$	$+$
f	$\nearrow$	$\searrow$	$\nearrow$

Clearly, the point $x = -4$ is a local maximum point, and the point $x = 2$ is a local minimum point.

Example 9.3. We want to construct a cylindrical can of maximum volume, which can be contained in a fixed conic container. We formulate the problem by setting R to be the base radius of the cone and h its height.

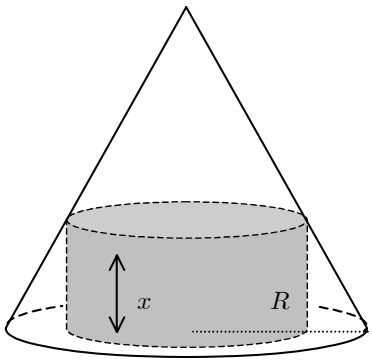


Figure 5.12. Cylinder inscribed in the cone of base radius R and height h

Every cylinder inscribed in the cone is univocally determined by its height, which we denote by x. We have to choose x in $(0, h)$, such that the volume of the cylinder is maximum. Now, the base radius r of the cylinder is[11] $r = R - \dfrac{R}{h}x$ and its volume is

$$V(x) = \pi \cdot \left(R - \frac{R}{h}x\right)^2 \cdot x = \pi \frac{R^2}{h^2} x(h-x)^2, \quad 0 < x < h.$$

We have

$$V'(x) = \pi \frac{R^2}{h^2}[(h-x)^2 - 2x(h-x)] = \pi \frac{R^2}{h^2}(h-x)(h-3x)$$

and therefore V' is zero at $x = h$ and $x = h/3$. Since $V(h) = 0$, corresponding to a degenerate cylinder (the segment forming the height of the cone), we expect that $x = h/3$ is the maximum point (actually a global maximum point). Indeed $V'(x) > 0$ for $0 < x < h/3$ and $V'(x) < 0$ for $h/3 < x < h$ and the conclusion follows from our test.

The required volume is therefore $V\left(\dfrac{h}{3}\right) = \dfrac{4}{27}\pi h R^2$.

5.10 De l'Hospital's theorem

The marquis Guillaume de l'Hospital (1661-1704) is popular among mathematicians because he wrote the first treatise on infinitesimal calculus. His name is linked to

[11] Use the proportion law $r : (h - x) = R : h$.

the theorem we are going to present now (the theorem was actually proved by Jean Bernoulli[12]). Its simplest version is often helpful for the computation of limits which present the indeterminate forms $0/0$ or ∞/∞.

To understand the basic idea of the theorem, let us consider the case $0/0$, that is we want to compute a limit of the type

$$\lim_{x\to c}\frac{f(x)}{g(x)}$$

where f and g tend to zero for $x \to c$:

$$\lim_{x\to c} f(x) = f(c) = 0, \qquad \lim_{x\to c} g(x) = g(c) = 0.$$

In such a situation the quotient between the two functions can be written as a quotient of... difference quotients: this suggests that there is a relation between this quotient and the quotient of the derivatives. We have indeed

$$\frac{f(x)}{g(x)} = \frac{f(x)-f(c)}{g(x)-g(c)} = \frac{f(x)-f(c)}{x-c} : \frac{g(x)-g(c)}{x-c}.$$

We pass to the limit for $x \to c$. If f and g are differentiable at c, we have

$$\lim_{x\to c}\frac{f(x)}{g(x)} = \frac{f'(c)}{g'(c)}$$

which is already a significant formula, at least when both derivatives are not zero. If we further suppose that the derivatives are continuous at c, we can write

$$\lim_{x\to c}\frac{f(x)}{g(x)} = \lim_{x\to c}\frac{f'(x)}{g'(x)}.$$

Conclusion: If the limit of the ratio between two functions is in the indeterminate form $0/0$, and we cannot compute it, then we can try to compute the limit of the ratio between the derivatives.

Actually, there is a subtlety hidden in the above description, which may cause problems: the two ratios (that of the functions and that of the derivatives) do not always have the same limit!

Oddly enough, it is the ratio of the derivatives which, usually, has an awkward behaviour. But, if this ratio "behaves well" (that is, it tends to a limit), then the other ratio is also forced to behave well and tend to the same limit. In rigorous terms, we finally have

Theorem 10.1 (de l'Hospital's theorem). *Let f and g be defined in a neighbourhood[13] U of c (c can also be $\pm\infty$) and satisfy the following conditions:*

[12] Jean Bernoulli (1667-1748), member of a family of mathematicians and scientists, had Euler as a student.

[13] This may also be only a left or right neighbourhood. In this case the limits are meant for $x \to c^-$ or $x \to c^+$.

(*i*) $\lim_{x\to c} f(x) = \lim_{x\to c} g(x) = 0$ *or* $\lim_{x\to c} f(x) = \lim_{x\to c} g(x) = \pm\infty$;
(*ii*) f *and* g *are differentiable in* U, *with* $g'(x) \neq 0$ *if* $x \neq c$;
(*iii*) $l = \lim_{x\to c} \dfrac{f'(x)}{g'(x)}$ *exists (finite or infinite).*
Then also

$$\lim_{x\to c} \frac{f(x)}{g(x)} = l.$$

Examples

10.1. We compute

$$\lim_{x\to 0} \frac{x - \sin x}{x^3}$$

The limit leads to the indeterminate form $0/0$. We try to apply de l'Hospital's theorem. The functions $f(x) = x - \sin x$ and $g(x) = x^3$ are differentiable in $\mathbb{R}$ and $g'(x) = 3x^2$ is zero only at $x = 0$. We then have

$$\lim_{x\to 0} \frac{f'(x)}{g'(x)} = \lim_{x\to 0} \frac{1 - \cos x}{3x^2}$$

which again leads to the indeterminate form $0/0$. Nothing forbids us from reapplying the theorem to the derivatives of $1 - \cos x$ and $3x^2$. We find:

$$\lim_{x\to 0} \frac{\sin x}{6x} = \frac{1}{6}.$$

All assumptions of de l'Hospital's theorem are satisfied: therefore we can conclude that

$$\lim_{x\to 0} \frac{x - \sin x}{x^3} = \frac{1}{6}.$$

10.2. We compute $\lim_{x\to 0^+} x^\alpha \ln x$, $\alpha > 0$, which leads to the indeterminate form $0 \cdot \infty$. We rewrite it as

$$\lim_{x\to 0^+} \frac{\ln x}{x^{-\alpha}},$$

which now leads to the indeterminate form ∞/∞. The functions $f(x) = \ln x$ and $g(x) = x^{-\alpha}$ are differentiable in $(0, +\infty)$ and $g'(x) = -\alpha x^{-\alpha-1}$ is never equal to zero in $(0, +\infty)$. We have

$$\lim_{x\to 0^+} \frac{f'(x)}{g'(x)} = \lim_{x\to 0^+} \frac{\dfrac{1}{x}}{-\alpha x^{-\alpha-1}} = \lim_{x\to 0^+} -\frac{1}{\alpha} x^\alpha = 0.$$

Since all assumptions of de l'Hospital's theorem are satisfied, we deduce that for $\alpha > 0$

$$\boxed{\lim_{x\to 0^+} x^\alpha \ln x = 0.}$$

In order to dissuade the reader from misusing the theorem, we give two examples in which the theorem shows its... limits.

10.3. We consider

$$\lim_{x\to+\infty} \frac{x+\sin x}{x},$$

of the form ∞/∞. We have differentiable functions in all $\mathbb{R}$ and the derivative of the denominator is 1 ($\neq 0$). We try and compute the limit of the ratio of the derivatives. We have

$$\lim_{x\to+\infty} (1+\cos x),$$

which does not exist. What can we deduce? Nothing, since the assumptions of de l'Hospital's theorem are not satisfied.

Nevertheless, without applying the limit, we have immediately:

$$\lim_{x\to+\infty} \frac{x+\sin x}{x} = \lim_{x\to+\infty} \frac{x}{x} = 1,$$

since $\sin x$, which varies between -1 and 1, is negligible w.r.t. x as $x \to +\infty$. This is a typical case in which the ratio of the derivatives has a behaviour which is clearly more irregular than the ratio of the functions.

10.4 (*Never ending...*). We consider

$$\lim_{x\to+\infty} \frac{\sqrt{x+1}}{\sqrt{x-1}} \tag{5.36}$$

which again leads to the indeterminate form ∞/∞, and we try to apply de l'Hospital's theorem. We easily see that the first two assumptions are satisfied. We then have, for the ratio of the derivatives,

$$\lim_{x\to+\infty} \frac{\dfrac{1}{2\sqrt{x+1}}}{\dfrac{1}{2\sqrt{x-1}}} = \lim_{x\to+\infty} \frac{\sqrt{x-1}}{\sqrt{x+1}}$$

which is very similar to the limit (5.36). What can we deduce? Nothing is wrong, but the formula does not help us to compute the limit and we have to calculate this limit using another method. Since

$$\sqrt{x-1} \sim \sqrt{x} \quad \text{and} \quad \sqrt{x+1} \sim \sqrt{x} \qquad \text{as} \quad x \to +\infty,$$

we deduce that the limit is equal to 1.

5.11 Taylor's formula

If f is differentiable at x_0, then its increment $f(x)-f(x_0)$ is well approximated, when x is near to x_0, by the differential $f'(x_0)(x-x_0)$. Sometimes, this linear approximation (also called the *first order approximation*) does not give enough information.

Think, for example, of a maximum problem for a function f. According to Fermat's Theorem, this problem can be (nearly) solved by finding the zeros of the derivative

f', which is equivalent to the condition that the graph has a horizontal tangent line. But this condition is necessary but not sufficient: it also holds true for the minimum points - which are exactly the opposite of what we are looking for. The tangent line "linearizes the graph" near the tangency point and cannot capture the difference between a maximum and a minimum (figure 13), which lies in the different way the two graphs are "non linear".

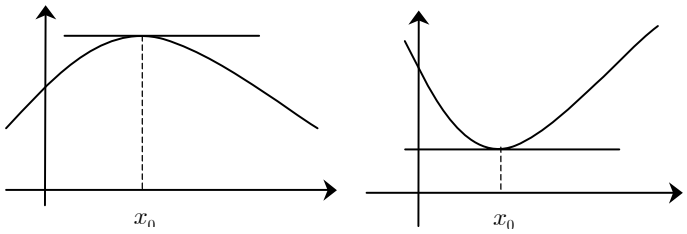


Figure 5.13. Stationary points of different nature

To better analyse the graph, we can try and estimate the difference between the increment $\Delta f = f(x) - f(x_0)$ of the function and its linearization, given by the differential $f'(x_0)(x - x_0)$. This difference

$$f(x) - [f(x_0) + f'(x_0)(x - x_0)] \tag{5.37}$$

represents the vertical difference betweeen the graph of f and its tangent line at the point of coordinates $(x_0, f(x_0))$.

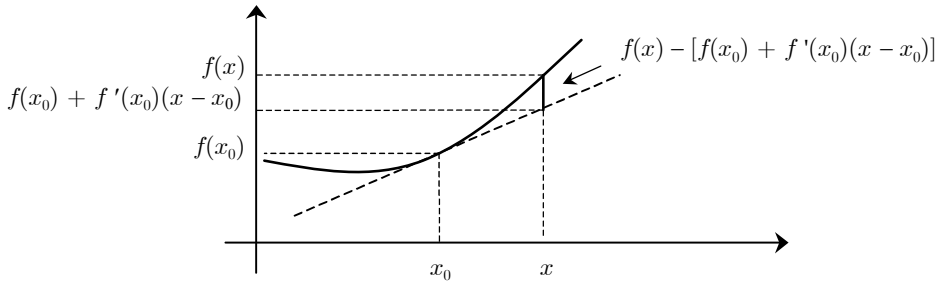


Figure 5.14. Difference between the graph of f and its tangent line

How can we *measure* therefore the distance of a graph from its linear approximation?

We can take the graph of a smooth function f and try to approximate it near the point x_0 with a curve which can follow the graph better than a straight line. The tangent line at $(x_0, f(x_0))$ has equation

$$y = f(x_0) + f'(x_0)(x - x_0).$$

To do better, we can try a parabola of equation

$$g(x) = f(x_0) + f'(x_0)(x - x_0) + a(x - x_0)^2,$$

in which we shall fix the parameter a later on. Now we observe that the parabola passes through the point $(x_0, f(x_0))$, since the value of g obtained for $x = x_0$ is exactly $f(x_0)$. The tangent to the parabola at x has slope

$$g'(x) = f'(x_0) + 2a(x - x_0),$$

which at x_0 is precisely $f'(x_0)$, the same as that of the graph of f. Therefore, the graph of f and the parabola share both the y-value and the tangent line at x_0.

The non-linear part of f translates into the variation of f', the slope of its graph. If we want a good approximation of f near x_0, a good idea is to look for the parabola whose slope varies exactly like that of the graph of f at the point x_0. Since the variation of a function is expressed by its derivative, we can ask that, at the point x_0, an equality holds not only between f' and g', but also between their derivatives.

We can define this second-born (the "derivative of the derivative") right away:

Definition 11.1. *If the derivative f' of f is differentiable, its derivative is called the* ***second derivative*** *of f and is denoted by one of the symbols*

$$f'', \qquad \frac{\mathrm{d}^2 f}{\mathrm{d}x^2}, \qquad \mathrm{D}^2 f, \qquad \ddot{f}.$$

As an example, we compute the second derivative of the function

$$f(x) = x^\alpha, \quad x > 0, \quad \alpha \in \mathbb{R}.$$

We have

$$f'(x) = \alpha x^{\alpha-1}, \quad f''(x) = \alpha(\alpha - 1)x^{\alpha-2}.$$

Let us go back to our problem, using these new symbols to formalize the requirement that the smooth function f and the quadratic function g are "similar" at x_0. We require that

$$f''(x_0) = g''(x_0).$$

Since

$$g''(x) = 2a,$$

we obtain

$$f''(x_0) = 2a,$$

and thus

$$a = \frac{1}{2} f''(x_0).$$

Therefore we have good hopes of approximating f near x_0 with the polynomial

$$T_2(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2$$

which is called the *Taylor's polynomial*[14] of the second order. What is interesting is the quality of the approximation, that is the order, as an infinitesimal, of the remainder

$$R(x) = f(x) - g(x) = f(x) - \left[f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2 \right],$$

i.e., the difference between the value of the function and of the quadratic polynomial T_2 at x. We have indeed

$$R(x) = o\left[(x - x_0)^2\right] \qquad \text{for } x \to x_0,$$

according to the following very important theorem.

Theorem 11.1. *If f is twice differentiable at x_0, then for $x \to x_0$*

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2 + o\left[(x - x_0)^2\right]. \tag{5.38}$$

The formula (5.38) is called *Taylor's formula up to the second order centred at x_0 with Peano's remainder*. In the particular case where $x_0 = 0$, the formula is called *Maclaurin's formula*[15] (still with Peano's remainder).

Proof. Formula (5.38) is equivalent to stating that

$$\frac{R(x)}{(x - x_0)^2} \to 0, \qquad \text{for } x \to x_0.$$

We have

$$\frac{R(x)}{(x - x_0)^2} = \frac{f(x) - f(x_0) - f'(x_0)(x - x_0) - f''(x_0)\dfrac{(x - x_0)^2}{2}}{(x - x_0)^2},$$

which takes the indeterminate form $0/0$ as x tends to x_0. We can see that the conditions required by de l'Hospital's theorem are satisfied. The quotient of the derivatives is

$$\frac{f'(x) - f'(x_0) - f''(x_0)(x - x_0)}{2(x - x_0)} = \frac{1}{2}\left[\frac{f'(x) - f'(x_0)}{x - x_0} - f''(x_0) \right].$$

The fraction in the square brackets is the difference quotient of f' and therefore tends to $f''(x_0)$ for $x \to x_0$. The quotient of the derivatives tends therefore to 0. Applying de l'Hospital's theorem, the quotient we started with also tends to 0 and the proof is complete. □

We write Maclaurin's formulae up to the second order in some important cases.

[14] Brook Taylor (1685-1731), English mathematician.
[15] Colin Maclaurin (1698-1746), Scottish mathematician.

Examples

11.1. If $f(x) = e^x$, we have $f(0) = f'(0) = f''(0) = 1$ and therefore

$$\boxed{e^x = 1 + x + \frac{x^2}{2} + o\left(x^2\right).}$$

11.2. Let $f(x) = \cos x$; we have $f'(x) = -\sin x$, $f''(x) = -\cos x$ and thus $f(0) = 1$, $f'(0) = 0$, $f''(0) = -1$. Therefore the formula is

$$\boxed{\cos x = 1 - \frac{x^2}{2} + o(x^2).}$$

11.3. Let $f(x) = \ln(1 + x)$. We have

$$f'(x) = \frac{1}{1+x}, \quad f''(x) = -\frac{1}{(1+x)^2}.$$

Thus $f(0) = 0$, $f'(0) = 1$, $f''(0) = -1$ and the formula is

$$\boxed{\ln(1 + x) = x - \frac{x^2}{2} + o(x^2).}$$

11.4. Let $f(x) = (1 + x)^\alpha$, α real. We have

$$f'(x) = \alpha(1 + x)^{\alpha-1}, \quad f''(x) = \alpha(\alpha - 1)(1 + x)^{\alpha-2}.$$

Therefore $f(0) = 1$, $f'(0) = \alpha$, $f''(0) = \alpha(\alpha - 1)$ and the formula is

$$\boxed{(1 + x)^\alpha = 1 + \alpha x + \frac{\alpha(\alpha - 1)}{2}x^2 + o\left(x^2\right).}$$

In particular, for $\alpha = 1/2$, we have

$$\boxed{\sqrt{1 + x} = 1 + \frac{x}{2} - \frac{x^2}{8} + o(x^2).}$$

In Figure 15 we show the graphs of the functions in examples 1-4 (dotted line), with their respective approximating parabolae (full line).

Second test for stationary points

Taylor's formula with Peano's remainder allows us to construct a test for recognizing the nature of stationary points, using the sign of the second derivative. In practice, we substitute the graph of f with the approximating parabola and we read from this parabola the nature of the point under consideration.

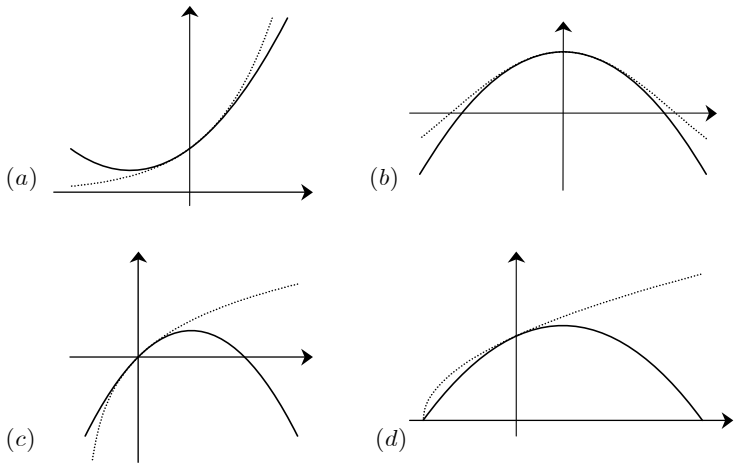


Figure 5.15. Graphs referring to examples 11.1, 11.2, 11.3 and 11.4

Theorem 11.2. *Let f be defined in (a,b), twice differentiable at x_0. Suppose further that $f'(x_0)=0$. If*

$$\begin{cases} f''(x_0)>0 & \text{then } x_0 \text{ is a point of (strict) local minimum} \\ f''(x_0)<0 & \text{then } x_0 \text{ is a point of (strict) local maximum.} \end{cases}$$

If x_0 is a stationary point for the function f, that is if $f'(x_0)=0$, we can write formula (5.38) in the form

$$f(x)-f(x_0)=\frac{f''(x_0)}{2}(x-x_0)^2+o\left[(x-x_0)^2\right].$$

This shows that the sign of the increment $f(x)-f(x_0)$ is given, for x very near to x_0, by the term $\dfrac{f''(x_0)}{2}(x-x_0)^2$, the other term being irrelevant. Now, if $f''(x_0)>0$ the sign of the increment will be positive, that is $f(x_0)$ is a local minimum, while if $f''(x_0)<0$ the sign of the increment will be negative, that is $f(x_0)$ is a local maximum.

Example 11.5. The function $f(x)=x\sin x$ has a stationary point at $x=0$. Indeed we have

$$f'(x)=\sin x+x\cos x \quad \text{and thus} \quad f'(0)=0+0\cdot 1=0.$$

Since

$$f''(x)=2\cos x-x\sin x$$

we have $f''(0)=2$. We can use test 11.2 and conclude that $x=0$ is a point of local minimum.

Remark. If $f''(x_0)=0$ the test cannot be used: the point x_0 may or may not be a point of local extremum. As examples, it is enough to consider the three functions

$$f_1(x)=x^4, \quad f_2(x)=-x^4, \quad f_2(x)=x^3.$$

All three have their first and second derivatives equal to zero at $x_0 = 0$, but for f_1 the origin is a minimum point, for f_2 it is a maximum one, while for f_3 it is neither a minimum point nor a maximum one.

The fact that the test using derivatives, in these cases, does not give any information is not surprising. The functions f_1, f_2, f_3 are powers of order higher than two, and therefore cannot be approximated, near to the origin, with a parabola, which is a power of order 2. The test can however be refined by using derivatives of higher order (see section 5.13).

Higher order derivatives

We have just seen here are good reasons for not stopping at the second derivative and for computing some other ones. If f'' is differentiable, its derivative is called the third derivative, and so on. The derivatives of order n (n natural integer) are denoted by the symbols

$$f^{(n)}, \qquad \frac{\mathrm{d}^n f}{\mathrm{d}x^n}, \qquad \mathrm{D}^n f.$$

We sometimes say that the derivative of *order zero* of f is f itself.

In the case of the power functions $f(x) = x^\alpha$, we have

$$f^{(n)}(x) = \alpha(\alpha-1)\cdots(\alpha-n+1)x^{\alpha-n}, \quad \text{for } \ n \geq 1.$$

For $\alpha = k$, a natural integer, we have

$$f^{(k)}(x) = k!, \qquad f^{(n)}(x) = 0 \quad \text{for } n > k.$$

5.12 Test for convexity (or concavity)

We have given a geometric definition of the notion of a *convex* function f in an interval $I \subseteq \mathbb{R}$, requiring that the part of the plane above its graph (that is its *epigraph*) is convex in the sense of elementary geometry. This is equivalent to requiring that the segment which joins any two points of the graph of f lies above that graph or partially coincides with the graph. The graph of a convex function may therefore contain some linear segments. When this does not happen, the function is called *strictly convex*. A function is *concave* or *strictly concave* if $-f$ is convex or strictly convex respectively.

This type of function appears very often in Economics; typically, the *consumer utility* function and the production function for factories are concave. It is helpful therefore to have some quick criteria for recognizing if a given function is convex (concave) or not, without having to verify directly if the definition holds.

We analyse the graph of a function f which is convex and smooth, such as those in Figure 16. We immediately find two fundamental features:

(a) proceeding from left to right, the slope of the graph always increases, or it remains constant (in the straight pieces of the graph).

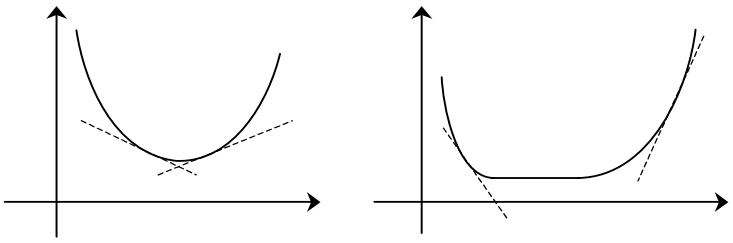


Figure 5.16. A strictly convex function and a convex (but not strictly convex) one

(b) if we plot the tangent line to the graph at any point, this line always lies below the graph, at most intersecting the graph along a segment and not only at a single tangency point.

Let us translate all of this into formulae. The first remark is equivalent to stating that f' is *monotonic increasing*. If f is twice differentiable, this latter statement is equivalent to the condition $f'' \geq 0$.

We therefore have our first method for recognizing the convexity (or concavity) of a function, by studying the monotonicity of its first derivative or the sign of its second derivative. We sum everything up in the following theorem.

Theorem 12.1 (Convexity test). *Let f be twice differentiable in an interval (a, b). Then the following three conditions are equivalent:*

(i) *f is convex*

(ii) *f' is increasing*

(iii) $f'' \geq 0$.

Of course, for *concave* functions, f' is *decreasing* and $f'' \leq 0$.

Example 12.1. We consider the function[16] $f(x) = x \ln x$, for $x > 0$. We have

$$f'(x) = \ln x + 1 \qquad f''(x) = \frac{1}{x}.$$

The (in this case *strict*) convexity of f can be deduced from the fact that f' is strictly increasing or from the fact that f'' is positive.

We now consider remark (b). The tangent line to the graph of f at a point x_0 has equation

$$y = f(x_0) + f'(x_0)(x - x_0).$$

If f is convex, its graph never goes below that line; this means that

$$f(x) \geq f(x_0) + f'(x_0)(x - x_0)$$

holds for every $x, x_0 \in (a, b)$.

[16] This function appears in some statistical indices (entropy) which have implications in Economics.

Theorem 12.2. *If f is differentiable in (a,b), the following conditions are equivalent:*

(i) *f is convex*

(ii) $f(x) \geq f(x_0) + f'(x_0)(x - x_0)$, *for every* $x, x_0 \in (a,b)$.

For concave functions it is enough to reverse the inequality in condition (ii). That condition also implies that every stationary point of a convex or concave function is a point of global minimum or maximum respectively.

Corollary 12.3. *If x_0 is a stationary point for a function f, convex (concave) in (a,b), then $f(x_0)$ is the minimum (maximum) value of f in (a,b).*

Proof. If f is convex and $f'(x_0) = 0$, condition (ii) implies that $f(x) \geq f(x_0)$ for every x in (a,b). For a concave f the opposite inequality sign holds. □

It is this property which makes convex functions extremely useful in optimization theory and in economical applications.

• *Optimal management of stocks.* A factory needs S tons of some goods for its annual production. Those goods are ordered at regular time intervals with quantity x, so that the annual number of orders is S/x. Every order has a fixed cost g, which is independent of the amount ordered. If we suppose that the velocity in using the goods in production is constant in time, and that there are no delays in the delivery of ordered goods, the average amount of goods in stock lies in the middle between the peak x, immediately after a delivery, and 0, immediately before remaining without goods. This average amount is therefore $(x+0)/2 = x/2$, equal to half of the ordered amount. Let us further suppose that to stock the goods costs m Euro per ton per year. The total inventory cost is

$$C(x) = g\frac{S}{x} + m\frac{x}{2}.$$

We are interested in finding the amount x^* which minimizes this cost. The derivative of the total cost is

$$C'(x) = -\frac{Sg}{x^2} + \frac{m}{2}$$

and is null at $x^* = \sqrt{2Sg/m}$. The second derivative is

$$C''(x) = \frac{2Sg}{x^3},$$

and is positive for every positive x. The second test for stationary points guarantees that x^* is a local minimum point, but the constant sign of C'' ensures that it is also a global minimum point.

• *Maximum profit.* A chemical factory would sell x tons of the washing powder **Sbiancador**, if the factory applied the unit price $p(x) = A/x^k$, with $A, k > 0$ and known from a market survey. The production cost of $x > 0$ tons of **Sbiancador** is the sum of the fixed cost F and the variable cost vx (with $v > 0$). Therefore the profit is

$$\pi(x) = xp(x) - F - vx = \frac{A}{x^{k-1}} - F - vx.$$

The factory is obviously interested in the production volume x^* which maximizes the profit. The derivative of the profit is

$$\pi'(x) = (1-k)\frac{A}{x^k} - v.$$

If $k \geq 1$, it is everywhere negative and it would be better to close the factory immediately. If $k < 1$, the derivative is zero at

$$x^* = \left[\frac{A(1-k)}{v}\right]^{\frac{1}{k}}.$$

To understand if this stationary point is a point of maximum, we compute π''; we have

$$\pi''(x) = -\frac{k(1-k)A}{x^{k+1}} < 0$$

for every $x > 0$. Since $\pi'' < 0$, π is concave and we deduce that x^* is a point of global maximum.

Points of inflection

We conclude with some considerations on inflection points. If we consider moving along the x-axis and meeting an inflection point x_0, we notice the change of the concavity of the function: from concave to convex or from convex to concave. Among other consequences, this implies that the tangent to the graph of f at the point $(x_0, f(x_0))$ has to "cross" the graph itself. It is therefore intuitive that, if f has a second derivative, it cannot be negative or positive at x_0. We therefore obtain:

Proposition 12.4. *If x_0 is an inflection point for f and if $f''(x_0)$ exists, then*

$$f''(x_0) = 0. \tag{5.39}$$

The condition that the second derivative of a function f is zero at a point x_0, expressed by the above theorem, does not automatically imply that x_0 is an inflection point. For example, the function $f(x) = x^4$ has second derivative equal to zero at $x = 0$ and nevertheless this point cannot be an inflection point, since f is convex in all of $\mathbb{R}$. An obvious way to find out if a point where formula (5.39) holds is an inflection point, is to study the sign of $f''(x)$ to the left and to the right of x_0: if the two signs are different it is an inflection point, otherwise it is not.

Example 12.2. The function

$$f(x) = x^4(x-1)^3$$

has second derivative

$$f''(x) = 6x^2(x-1)\left[2(x-1)^2 + 4x(x-1) + x^2\right]$$

which is null both at $x = 0$ and at $x = 1$. We notice that, near those points, the expression within square brackets is nearly equal to, respectively, 2 and 1. Therefore

this expression has no influence on the sign of $f''(x)$. The factor which annihilates the second derivative at 0 is x^2, which *does not change sign* in a neighbourhood of 0. Therefore, the concavity does not change. This is completely different from what happens for $(x-1)$, the factor which is responsible for the other zero of $f''(x)$. Therefore the origin is not an inflection point, while $x=1$ is. This is also confirmed by the graph in Figure 17.

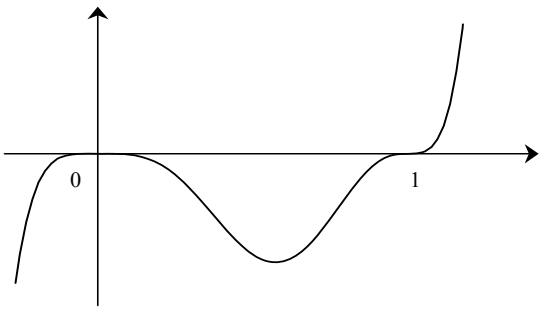


Figure 5.17. Graph of $f(x) = x^4(x-1)^3$

5.13 Taylor's formula of order n

Taylor's formula with Peano's remainder

We want to generalize the formula we introduced in section 11, and approximate a function f with a polynomial of degree $n \geq 2$, near to a point x_0. We consider a point x near to x_0, let h be the increment of x w.r.t. x_0: our goal is to express $f(x_0+h)$ as the sum of a polynomial $T_n(h)$ (of degree n at most) and a negligible error term. How can we determine $T_n(h)$?

The underlying idea is that the polynomial that best approximates f near x_0 must have all derivatives (computed for $h=0$) equal to the derivatives of f at x_0. Let us now see how $T_n(h)$ should appear. Setting

$$T_n(h) = a_0 + a_1 h + a_2 h^2 + \cdots + a_n h^n,$$

its derivatives are:

$$\begin{aligned} T_n'(h) &= a_1 + 2a_2 h + \cdots + n a_n h^{n-1}, \\ T_n''(h) &= 2a_2 + \cdots + n(n-1) a_n h^{n-2}, \\ &\vdots \\ T_n^{(n)}(h) &= n! a_n. \end{aligned}$$

Substituting $h=0$, we obtain the conditions

$$T_n(0) = a_0, \quad T_n'(0) = a_1, \quad T_n''(0) = 2a_2, \ \ldots\ , \ T_n^{(n)}(0) = n! a_n,$$

from which we deduce

$$a_0 = f(x_0), \quad a_1 = f'(x_0), \quad 2a_2 = f''(x_0), \ \ldots, \ \ n!a_n = f^{(n)}(x_0),$$

therefore the coefficient of the k-th term (for $k = 0, \ldots, n$) is

$$a_k = \frac{f^{(k)}(x_0)}{k!}.$$

Our candidate for the Taylor polynomial of order n[17] is therefore

$$T_n(h) = f(x_0) + f'(x_0)h + \frac{f''(x_0)}{2}h^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}h^n.$$

In what sense does $T_n(h)$ approximate f well near x_0? The following theorem tells us that, if we substitute the value $f(x)$ at a point $x = x_0 + h$ near to x_0 with the value $T_n(h)$, we create an error which is negligible w.r.t. h^n (i.e., as h tends to 0 the error tends to 0 faster than h^n).

Theorem 13.1. Taylor's formula (with Peano's remainder).
If $f : (a, b) \to \mathbb{R}$ is n-times differentiable at $x_0 \in (a, b)$, the following formula holds:

$$f(x_0 + h) = f(x_0) + f'(x_0)h + \frac{f''(x_0)}{2}h^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}h^n + o(h^n), \quad \text{for } h \to 0.$$

By setting $f^{(0)}(x_0) = f(x_0)$, the above formula can be rewritten as

$$f(x_0 + h) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}h^k + o(h^n), \quad \text{for } h \to 0. \tag{5.40}$$

Formula (5.40) is called the *Taylor's formula of order n, centred at x_0, with Peano's remainder*. In terms of $x = x_0 + h$, formula (5.40) becomes

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + o\left[(x - x_0)^n\right], \quad \text{for } x \to x_0.$$

We omit the proof.

• *Maclaurin's formula.* In the particular case in which $x_0 = 0$, we obtain the so-called Maclaurin's formula (with Peano's remainder):

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2}x^2 + \cdots + \frac{f^{(n)}(0)}{n!}x^n + o(x^n), \quad \text{for } x \to 0.$$

We now list Maclaurin's formulae (of order n with Peano's remainder) for the elementary functions.

[17] Obviously, its degree is n only if $f^{(n)}(x_0) \neq 0$.

13.1. Let $f(x) = e^x$. Since $f^{(n)}(x) = e^x$ and $f^{(n)}(0) = 1$ for all $n \geq 1$, we have

$$e^x = 1 + x + \frac{x^2}{2} + \cdots + \frac{x^n}{n!} + o(x^n), \quad \text{for } x \to 0.$$

The graph of the function $f(x) = e^x$ and its approximating polynomials of degree 3 and 6 are plotted in Figure 18 (a).

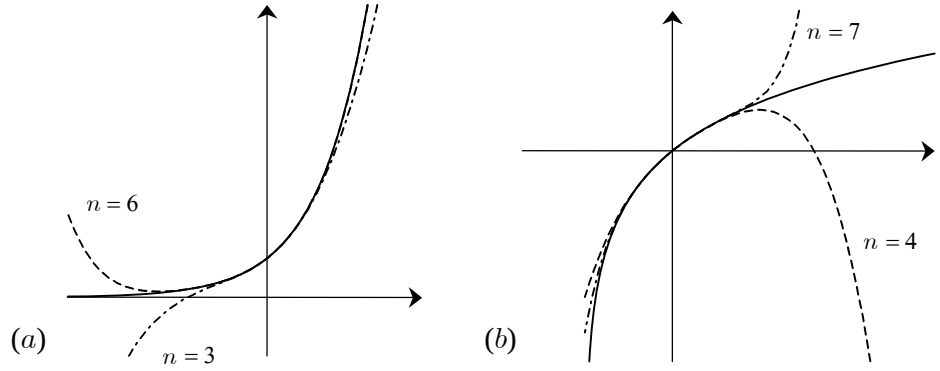


Figure 5.18. Graphs of the functions $f(x) = e^x$ and $f(x) = \ln(1+x)$ with some approximating polynomials

13.2. Let $f(x) = \ln(1+x)$. We have

$$\begin{aligned} f'(x) &= \frac{1}{1+x}, \quad f''(x) = -\frac{1}{(1+x)^2}, \\ f'''(x) &= \frac{2}{(1+x)^3}, \ \ldots \ , \ f^{(n)}(x) = (-1)^{n-1}\,\frac{(n-1)!}{(1+x)^n} \end{aligned}$$

and therefore

$$\begin{aligned} f(0) &= 0, \quad f'(0) = 1, \quad f''(0) = -1, \\ f'''(0) &= 2, \ \ldots \ , \ f^{(n)}(0) = (-1)^{(n-1)}(n-1)!. \end{aligned}$$

The formula is

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} + \cdots + (-1)^{(n-1)}\frac{x^n}{n} + o(x^n), \quad \text{as } x \to 0.$$

The graph of the function $f(x) = \ln(1+x)$ and its approximating polynomials of degree 4 and 7 are plotted in Figure 18 (b).

13.3. Let $f(x) = \sin x$ and $g(x) = \cos x$. We have

$$\begin{aligned} f'(x) &= g(x) = \cos x, \quad f''(x) = g'(x) = -\sin x, \\ f'''(x) &= g''(x) = -\cos x, \quad f^{(iv)}(x) = g'''(x) = \sin x, \end{aligned}$$

and so on. At 0 the derivatives of even order of f and the derivatives of odd order of g are zero, the other ones take values $+1$ or -1 alternatively. We thus have

$$\begin{aligned}\sin x &= x - \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots + (-1)^n \frac{x^{2n+1}}{(2n+1)!} + o(x^{2n+2}), \quad \text{for } x \to 0,\\ \cos x &= 1 - \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots + (-1)^n \frac{x^{2n}}{(2n)!} + o(x^{2n+1}), \quad \text{for } x \to 0.\end{aligned}$$

The graph of the function $f(x) = \sin x$ and its approximating polynomials of degree 5 and 11 are plotted in Figure 19.

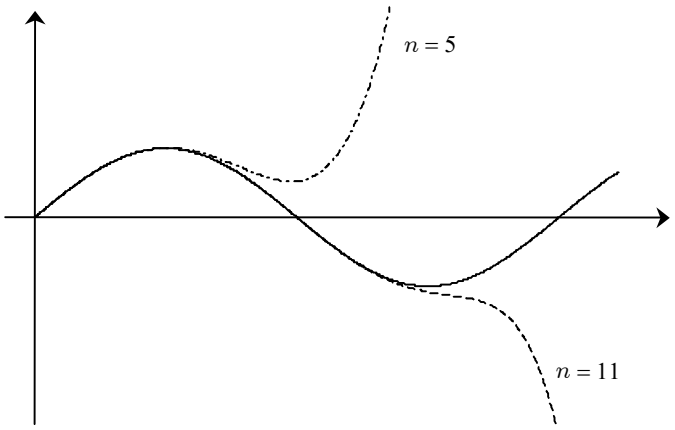


Figure 5.19. Graph of the function $f(x) = \sin x$ with some approximating polynomials

13.4. Let $f(x) = (1+x)^a, \quad \alpha \in \mathbb{R}$. We have

$$\begin{aligned}f'(x) &= \alpha\,(1+x)^{\alpha-1}\,, \quad f''(x) = \alpha\,(\alpha-1)\,(1+x)^{\alpha-2}\,, \ \ldots,\\ f^{(n)}(x) &= \alpha\,(\alpha-1)\cdots(\alpha-n+1)(1+x)^{\alpha-n},\end{aligned}$$

and therefore

$$\begin{aligned}f(0) &= 1, \quad f'(0) = \alpha, \quad f''(0) = \alpha\,(\alpha-1)\,, \ \ldots,\\ f^{(n)}(0) &= \alpha\,(\alpha-1)\cdots(\alpha-n+1).\end{aligned}$$

Setting

$$\binom{\alpha}{k} = \frac{\alpha\,(\alpha-1)\cdots(\alpha-k+1)}{k!},$$

for every real α and for every natural number k, we obtain the formula

$$(1+x)^\alpha = 1 + \alpha x + \binom{\alpha}{2} x^2 + \cdots + \binom{\alpha}{n} x^n + o(x^n), \quad \text{for } x \to 0.$$

In the case $\alpha = n$, the formula gives Newton's binomial theorem.

We want to stress the fact that Taylor's formula gives information about the local behaviour of a function: in a neighbourhood of x_0, f is well approximated by its Taylor's polynomials. The formula does not give information about the width of such a neighbourhood: the error we make when we approximate a function by its Taylor's polynomial of order n is marginal w.r.t. $(x - x_0)^n$. The graphs of the exponential function and the sine function may lead one to think that, given a fixed interval, if we increase the degree of Taylor's polynomial this will provide a better approximation of the function on the whole interval, but this is not generally true. And this is indeed confirmed by looking at the graph of the logarithmic function.

Test for stationary points

The following theorem generalizes the second test for stationary points. It is useful in unlucky situations, where quite a number of derivatives (after the first one) are equal to zero at the stationary point. The test simply consists in going on computing derivatives, until we find a non-zero one.

Theorem 13.2. *Let $f : (a, b) \to \mathbb{R}$ be n-times differentiable in $x_0 \in (a, b)$ with*

$$f'(x_0) = f''(x_0) = \cdots = f^{(n-1)}(x_0) = 0, \qquad f^{(n)}(x_0) \neq 0.$$

If n is even and

$$f^{(n)}(x_0) > 0$$

then x_0 is a point of (strong) local minimum; if n is even and

$$f^{(n)}(x_0) < 0$$

then x_0 is a point of (strong) local maximum; if n is odd, then x_0 is not a point of local extremum.

Proof. We must study, for small $|h|$, the sign of the increment

$$\Delta f = f(x_0 + h) - f(x_0).$$

We may use Taylor's formula, as the hypothesis of Theorem 13.1 is satisfied. We write

$$f(x_0 + h) = f(x_0) + \frac{f^{(n)}(x_0)}{n!} h^n + o(h^n), \quad \text{for } h \to 0,$$

and we deduce

$$f(x_0 + h) - f(x_0) = \frac{f^{(n)}(x_0)}{n!} h^n [1 + o(1)], \quad \text{for } h \to 0,$$

since the expression $\dfrac{n!}{f^{(n)}(x_0)} \dfrac{o(h^n)}{h^n}$ is $o(1)$ for $h \to 0$, by definition of the symbol "o".

Now, the expression between square brackets tends to 1 for $h \to 0$, thus it is positive in a neighbourhood of x_0 by the theorem on the permanence of sign. The sign of the increment Δf therefore depends on the sign of $f^{(n)}(x_0)$ and of h^n.

If n is even, h^n is positive for $h \neq 0$ and, thus, the sign of Δf agrees with the sign of $f^{(n)}(x_0)$: if $f^{(n)}(x_0) > 0$, then Δf is positive in a neighbourhood of x_0 ($x \neq x_0$) and x_0 is a point of strong local minimum; if $f^{(n)}(x_0) < 0$, then Δf is negative in a neighbourhood of x_0 ($x \neq x_0$) and x_0 is a point of strong local maximum.

If n is odd, the sign of h^n depends on h and thus the sign of Δf changes according to the sign of h. Therefore x_0 is not a point of local extremum. □

Taylor's formula with Lagrange's remainder

The formula we have seen in theorem 13.1 holds only locally, as we have stressed in various remarks. On the other hand, we have seen some examples (especially the one concerning the sine function) which lead us to think that, with a Taylor's polynomial of sufficiently high degree, the approximation may be "good" not only locally. We therefore ask ourselves if it is possible to approximate a function on a fixed interval with a Taylor's polynomial, and to have a precise estimate of the error we make.

Given a function $f : (a, b) \to \mathbb{R}$, differentiable at least $n + 1$ times, and a point $x_0 \in (a, b)$, we want to estimate the difference

$$f(x) - f(x_0) - f'(x_0)(x - x_0) - \frac{f''(x_0)}{2}(x - x_0)^2 + \cdots - \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$$

on the whole interval (a, b).

We start by considering $n = 1$ and trying to evaluate the difference

$$R(x) = f(x) - f(x_0) - f'(x_0)(x - x_0),$$

which represents the difference between f and its tangent line. We assume that such a difference is proportional to $(x - x_0)^2$ and that we can write

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + b(x - x_0)^2$$

in which we leave the coefficient b as unknown (the reader can already imagine that we shall try to choose the coefficient b in a clever way).

The theorem below allows us to find an upper bound for the error, providing a "static" representation of the remainder $R(x)$, where static means "given a fixed x".

Theorem 13.3. *Let f be a twice differentiable function in the interval (a, b) and let x_0, x be two points of (a, b). Then there exists at least one point c between x_0 and x such that*

$$\boxed{f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(c)}{2}(x - x_0)^2.} \tag{5.41}$$

Equation (5.41) is called *Taylor's formula of order two with centre at x_0 and Lagrange's remainder.*

Writing this formula as

$$\frac{f(x) - f(x_0)}{x - x_0} = f'(x_0) + \frac{f''(c)}{2}(x - x_0)$$

it clearly appears as a generalization of Lagrange's theorem. Indeed, its proof is based precisely on this theorem.

Proof. Setting

$$w(x) = f(x) - f(x_0) - f'(x_0)(x - x_0) - b(x - x_0)^2,$$

we get

$$w'(x) = f'(x) - f'(x_0) - 2b(x - x_0).$$

This derivative is zero at x_0 and we may choose b in such a way that also $w'(x) = 0$: indeed it suffices that

$$b = \frac{1}{2}\frac{f'(x) - f'(x_0)}{x - x_0}.$$

The hypotheses of Lagrange's theorem hold, and therefore there exists one point c between x_0 and x such that $w''(c) = 0$. Since

$$w''(x) = f''(x) - 2b,$$

there exists c between x_0 and x such that

$$f''(c) - 2b = 0,$$

and thus

$$b = \frac{f''(c)}{2}. \quad \square$$

From the expression we found for Lagrange's remainder, we may deduce a convenient upper bound for it. If the second derivative of f does not exceed in absolute value a number $M > 0$ when the argument of f varies between x_0 and x, we can assert that the approximation error does not exceed (in absolute value)

$$\frac{M}{2}(x - x_0)^2.$$

We consider, as an example, the function $\sin x$ near $x_0 = 0$. This new form of the remainder of Taylor's formula ensures that

$$\sin x = \sin 0 + x\cos 0 + \frac{1}{2}(-\sin c)\,x^2 = x + \frac{x^2}{2}(-\sin c),$$

with c between 0 and x. Since the sine function never exceeds 1 in absolute value ($M = 1$), we obtain

$$|R(x)| = |\sin x - x| \leq \frac{x^2}{2}$$

Therefore, by substituting $\sin 1/100$ with $1/100$, we would make an error which is not larger than $1/20000 = 0.00005$.

This Taylor's formula may be extended beyond the first order in an analogous way.

Theorem 13.4. **Taylor's formula (with Lagrange's remainder).** *Let $f : (a, b) \to \mathbb{R}$ be $(n+1)$-times differentiable in (a, b), and let $x_0,\ x \in (a, b)$. Then there exists at least one point c between x_0 and x such that*

$$f(x) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k + \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}. \tag{5.42}$$

By setting $x = x_0 + h$, formula (5.42) becomes

$$f(x_0 + h) = \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}h^k + \frac{f^{(n+1)}(c)}{(n+1)!}h^{n+1},$$

with c a suitable point between x_0 and $x_0 + h$.

Without using summation indices, formula (5.42) becomes:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1}.$$

Furthermore, if we succeed in proving that $f^{(n+1)}$ does not exceed in absolute value a certain constant M, i.e., if $|f^{(n+1)}(x)| \leq M$ for all $x \in (a, b)$, then also $|f^{(n+1)}(c)| \leq M$ and we obtain an estimate for the error obtained by substituting f with its Taylor's polynomial.

$$|f(x_0 + h) - T_n(h)| = \frac{|f^{(n+1)}(c)|}{(n+1)!}\,|h|^{n+1} \leq \frac{M}{(n+1)!}\,|h|^{n+1}\,.$$

• *Maclaurin's formula.* If we set $x_0 = 0$ in formula (5.42), we obtain the so-called Maclaurin's formula with Lagrange's remainder:

$$f(x) = f(0) + f'(0)x + \frac{f''(0)}{2}x^2 + \cdots + \frac{f^{(n)}(0)}{n!}x^n + \frac{f^{(n+1)}(c)}{(n+1)!}x^{n+1}. \tag{5.43}$$

5.14 Exercises

5.1. Compute the derivative of the following functions

$$\begin{array}{lll} (a)\ y = x^3 e^{5x}, & (b)\ y = (\sin x)^7\,, & (c)\ y = \dfrac{x}{1+x^2}, \\ (d)\ y = \ln\dfrac{1-x}{1+x}, & (e)\ y = \arctan\dfrac{1}{x}, & (f)\ y = \ln\left(x + \sqrt{1+x^2}\right). \end{array}$$

5.2. Following the steps in example 5.2, calculate the derivative of

$$y = x^x.$$

Generalize the result by writing a formula for the derivative of

$$y = f(x)^{g(x)}.$$

5.3. Let x be the income of a person and $f(x)$ the total tax amount. The difference quotient of f has the meaning of the average tax rate corresponding to the supplementary income of h Euro for a person with an income x_0, while the derivative of f is the so-called *marginal tax rate* (see formula (5.6)). What happens to f' at the points where the tax rate passes from one class to another, that is at points such as point a, for a function f of type

$$f(x) = \begin{cases} \alpha x & \text{for } x \leq a \\ \alpha a + \beta(x - a) & \text{for } x > a, \end{cases}$$

usually with $\beta > \alpha$ (the so-called *progressivity of the income tax*)?

5.4. (⇛ **Chapter 11**) Financial calculations show the existence of a link between the annual compound interest rate i and the equivalent force of interest $\delta = \ln(1 + i)$. Compare the value of δ with i on the basis of graphic considerations (compute and plot the tangent to the graph of δ for $i = 0$).

5.5. Compute the elasticity of a linear demand function

$$q(p) = a - bp.$$

5.6. (⇛ **Chapter 11**) Compute the *force of interest*

$$\delta(t) = D[\ln f(t)] = \frac{f'(t)}{f(t)}$$

for the three usual systems of interest laws:

$$\begin{array}{ll} (a)\ f(t) = 1 + it & \text{simple interests,} \\ (b)\ f(t) = (1 + i)^t & \text{compound interests,} \\ (c)\ f(t) = \dfrac{1}{1 - dt} & \text{anticipated simple interests.} \end{array}$$

5.7. The function

$$x(t) = 10000 \frac{e^t}{1 + e^t}$$

describes the growth of a population as a function of time (x = average number of people at time t).

Plot the graph of the function $x(t)$.

For which value of t do we obtain the maximum growth rate $x'(t)$?

5.8. The function

$$f(x) = \frac{x^2}{100} + \frac{x}{50} + 400, \quad x > 0$$

represents the *total cost* of production for a certain kind of goods, as a function of the produced quantity x.

Write the expression of the average cost $g(x) = f(x)/x$. Compute its minimum and plot a qualitative graph of g. Compute the elasticity of f and check that at the minimum point of g the elasticity of f is equal to 1.

5.9. The demand function for a certain kind of goods is

$$q(p) = 1500e^{-0.025p}$$

where p is the price.

Determine the price which maximizes the total revenue. Check that the elasticity of q is 1 when the total revenue is maximized.

5.10.(⇛ **Chapter 11**) The DCF of a financial operation, with cash flows at the maturities indicated in the table

maturity	**cash flow**
0	−1000
1/2	+100
1	+1100

is

$$G(x) = -1000 + \frac{100}{\sqrt{1+x}} + \frac{1100}{1+x}.$$

Plot the graph of the function G when the interest rate x varies in the domain $x > -1$.

5.11. To produce x units of some goods, a factory spends

$$C(x) = 1000 + 100x - 50x^2 + 10x^3.$$

The selling price of one unit of goods is 400.

(a) Construct the function $\pi(x)$, which expresses the profit for the factory as a function of x, the produced amount of the considered goods. Compute $\pi'(x)$ and find the values $x = x^*$ for which the profit is maximum.

(b) Compute the functions of marginal cost $C_{\text{ma}}(x)$ and average cost $C_{\text{av}}(x)$. Check that, where the derivative of the average cost is zero, its value coincides with the marginal cost.

5.12. A function which gives the quantity of some produced goods $f(x)$ as a function of the amount x of some employed factor (for example: the number of workers) may, in some cases, be convex before a point and concave after that point. At first (where it is convex) the marginal productivity of the factor increases when the amount of produced goods increases (these are the so-called Economies of Scale). Which assumptions would you require on $f'(x)$ and on $f''(x)$ to guarantee such a behaviour?

5.13. In some marketing models, one often assumes that the velocity of growth of the amount of sales $f(x)$ as a function of the advertising costs is directly proportional to the residual market $M - f(x)$, given by the difference between the potential market M and the conquered market $f(x)$. Explain why if a relation of type:

$$f'(x) = a[M - f(x)] \quad \text{with } a > 0$$

holds, then f is an increasing and concave function.

5.14. Logistic models are often used to describe the diffusion of new products on the market. They are based on functions of type:

$$V(t) = \frac{a}{1 + be^{-ct}}$$

where $a, c > 0$ and $b > 1$.

Compute the amount of sales for $t = 0$. Compute $\lim_{t \to +\infty} V(t)$. Prove that V is monotonic increasing. Show that its graph has an inflection point, which is called the time of maturity for the product. Compute the time of maturity for the product and the fraction of market already conquered at that time.

5.15. The value of a good wine increases with time. A model with a certain reputation assigns to a bottle of $x \geq 0$ years the value

$$v(x) = h\sqrt{1 + x/a},$$

where h is the initial value and a is a positive parameter. We denote by $a(x)$ the discounted value of the price of a bottle, calculated at a compound interest rate i. Determine $a'(x)$ and show that this discounted value at first increases and then decreases, according to the sign of a'.

Compute the optimal stock period x^*, that is the period for which the discounted price is maximum. Assuming that a bottle has a maximum discounted value for $x^* = 15$ years and that $i = 1\%$, deduce an estimate for the parameter a.

5.16. The value of a capital, invested in a forest, varies with time t according to the function

$$M(t) = bt^c e^{-kt},$$

with $b, c, k > 0$.

(a) Compute the derivative of this function and show that it is positive between 0 and t^*, and negative after t^*. Deduce the age of maximum value for the forest.

(b) Plot the graph of M in correspondence of the values $b = 10000$, $c = 2$, $k = 1$.

(c) Compute $\rho(t)$, the logarithmic derivative of $M(t)$, which can be interpreted as the force of interest of the forest investment.

5.17. A company's total inventory costs is made up of the cost $g = 100$ Euro for every issued order and the warehousing costs $m(z) = 80z$, which is a function of the average amount z of goods stored in the stock.

Let $S = 1000$ (tons) be the annual need of goods, which are regularly used in time, and x the amount of each order. Construct the function $C(x)$ which gives the total inventory costs and determine for what order quantity the inventory cost is minimum.

5.18. Compute the limits

$$\lim_{x \to 0} \frac{\sqrt[3]{27 + x} - 3}{x}, \qquad \lim_{x \to +\infty} \frac{\ln\left(3 + e^{4x}\right)}{x}.$$

5.19. Write the Taylor's polynomials of the second order which approximate the following functions in a neighbourhood of the given point.

$$\begin{aligned} f(x) &= \ln(3+x), \qquad x_0 = 0; \\ g(x) &= \frac{1}{x}, \qquad\qquad x_0 = 2. \end{aligned}$$

Then write Taylor's polynomials of the fourth order for the same functions (at the same points).

5.20. Consider the function (*Bernoulli's utility*)

$$u(x) = a \cdot x^b,$$

with $a > 0$ and $0 < b < 1$. Compute its first three derivatives. Compute the function

$$a(x) = -\frac{u''(x)}{u'(x)} \quad (\textit{risk aversion}).$$

In some parts of Economics it is also of interest to compute the function

$$p(x) = -\frac{u'''(x)}{u''(x)}.$$

Compute $p(x)$; $p(x)$ is called *prudence*.

5.21. A factory sells x units of a certain kind of goods on a market where the demanded quantity depends on the price p according to the isoelastic relation

$$x = a/p^k,$$

with $a > 0$ and $k > 1$. This factory has variable costs given by a total amount v per unit of product, to which the cost c of one purchased component has to be added (still per unit of product).

Construct the profit function $\pi(x)$ for the factory and find the value x^* for which this function is maximum.

The obtained answer may be interpreted as the demand function for the supplier of the component. What is the best way the supplier can choose c, so that his profit $g(c)$ is maximum, considered that he sustains a variable unitary cost h?

5.22. The total cost of some goods, as a function of the produced quantity, is given by the function $C = f(q)$, $q \in [a, b]$, with f differentiable and convex. Prove that if at the point q_0 the elasticity of f is equal to one, then q_0 is a point of minimum for the function

$$C_M = \frac{f(q)}{q}.$$

6

Series

The combination of the two operations of sum and limit leads to the definition of the concept of *series* or *infinite sum*. The idea itself is quite easy, and series are certainly a useful tool. Nevertheless, some difficult points require attention in order to avoid serious errors. The chapter is structured along the following lines.

• The definition of *series* is formalized and *series* are classified as *convergent, divergent*, or *irregular* according to their behaviour, as we have already done with sequences.

• The next point is to recognize the behaviour of a given series (convergent, divergent, or irregular). We deal separately with:
— criteria for series having terms of constant sign;
— criteria for series having terms with alternate sign.

6.1 The concept of series

We start with a "classical" example which reminds us of the famous Zeno's paradox. Suppose we want to measure the length of a 1–meter–long rod using the following procedure. We divide the rod into two halves and we measure the first piece: we obtain 1/2. Then we divide the second piece in half and we measure the first part: we obtain $1/2 \times 1/2 = 1/4$. Then we divide the second part in half and we continue this way. The different parts of the rod which we obtain are infinitely many: their lengths, arranged in decreasing order, form the (geometric) sequence

$$\left\{\frac{1}{2}, \frac{1}{2^2}, \frac{1}{2^3}, \ldots, \frac{1}{2^n}, \ldots\right\}.$$

We should expect that by summing all the lengths we would get the length of the whole rod. The mathematical formula corresponding to this conjecture is

$$\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+\ldots+\frac{1}{2^n}+\ldots=1. \tag{6.1}$$

The problem with (6.1) is that the left hand side is an "infinite sum": but what is the mathematical meaning of summing up infinitely many numbers?

A quite natural idea is the following: instead of considering the sum of *all* the terms, we consider the sum of the *first n terms*

$$S_n=\frac{1}{2}+\frac{1}{2^2}+\frac{1}{2^3}+\ldots+\frac{1}{2^n}=\sum_{k=1}^{n}\frac{1}{2^k},$$

and then we let n go to $+\infty$. We expect that

$$\lim_{n\to+\infty} S_n=1.$$

Indeed, recalling how S_n was constructed, we have

$$S_n=\ 1-\frac{1}{2^n},$$

and

$$\lim_{n\to+\infty} S_n=\lim_{n\to+\infty}\left(1-\frac{1}{2^n}\right)=1.$$

We can therefore write

$$\sum_{k=1}^{+\infty}\frac{1}{2^k}=1.$$

Generally speaking, we have:

Definition 1.1. *Given a sequence $\{a_0,a_1,\ldots,a_n,\ldots\}$ of real numbers, we define a* ***series*** *as the symbol*

$$\sum_{k=0}^{+\infty} a_k. \tag{6.2}$$

The terms of the sequence $\{a_n\}$ are called terms of the series.

With every *series* we associate the sequence $\{S_n\}$, called the sequence *of partial sums*, having general term [1]

$$S_n=\sum_{k=0}^{n} a_k.$$

The sequence $\{S_n\}$ may also be recursively defined by

$$\begin{cases} S_0=a_0 \\ S_n=S_{n-1}+a_n. \end{cases} \tag{6.3}$$

[1] S_n is the sum of all terms in $\{a_n\}$ *up to the n-th one*, whatever the initial index.

Exactly as with sequences, series are divided into three types: *convergent, divergent, and irregular*. To establish the *behaviour* of a series means to determine its type, precisely:

Definition 1.2 (behaviour of a series). *The series* (6.2) *is called:*

(*a*) ***convergent*** *to the* ***sum*** S, *if* S_n *converges to* S. *We write*

$$\boxed{\sum_{s=0}^{+\infty} a_s = S}$$

(*b*) ***divergent*** *to* $+\infty$ *or* $-\infty$, *if* S_n *diverges to* $+\infty$ *or* $-\infty$. *We write respectively*

$$\boxed{\sum_{s=0}^{+\infty} a_s = +\infty \quad or \quad \sum_{s=0}^{+\infty} a_s = -\infty}$$

(*c*) ***irregular*** *if the limit of* S_n *does not exist.*

In the cases (*a*) and (*b*) (convergent or divergent series) the series is said to be *regular*. We may also say that a series diverges to ∞ (without sign), meaning that $\lim |S_n| = +\infty$.

Example 1.2. Let us see a famous converging series, the so-called *Mengoli series*[2]:

$$\frac{1}{1\ \cdot\ 2} + \frac{1}{2\ \cdot\ 3} + \frac{1}{3\cdot\ 4} + \cdots + \frac{1}{k\,(k+1)} + \cdots = \sum_{k=1}^{+\infty} \frac{1}{k\,(k+1)}.$$

By observing that

$$\frac{1}{1\cdot 2} = 1 - \frac{1}{2}, \quad \frac{1}{2\cdot 3} = \frac{1}{2} - \frac{1}{3}, \ldots, \quad \frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1},$$

the sequence of its partial sums

$$S_n = \frac{1}{1\cdot 2} + \frac{1}{2\cdot\ 3} + \frac{1}{3\cdot\ 4} + \cdots + \frac{1}{n\,(n+1)}$$

may be rewritten as

$$S_n = \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \cdots + \left(\frac{1}{n} - \frac{1}{n+1}\right).$$

We see that its terms cancel out two by two and we have

$$S_n = 1 - \frac{1}{n+1} \to 1$$

[2] Pietro Mengoli (1625-1686).

Therefore, the Mengoli series converges and has sum 1.

The particularity of this series is that its general term a_n may be rewritten as the difference between two consecutive terms of a same sequence $\{b_n\}$ (i.e. $a_n = b_n - b_{n+1}$). This is the reason which allows us to compute the sum. In this case the sequence $\{b_n\}$ is $\{1/n\}$. Series of this type are called *telescopic*.

Remark 1.1. The behaviour and the sum of a series do not depend on the symbol which is used as index in the formula (6.2); indeed

$$\sum_{n=0}^{+\infty} a_n \quad \text{and} \quad \sum_{i=0}^{+\infty} a_i$$

have the same meaning.

Remark 1.2. The behaviour of a series is determined by the behaviour of the sequence of partial sums. This behaviour does not depend on the first n terms of the series, however large n is. Two series, which differ only by a finite number of terms, have the same behaviour. In particular

$$\sum_{n=0}^{+\infty} a_n \quad \text{and} \quad \sum_{n=k}^{+\infty} a_n$$

have the same behaviour, for every $k > 0$. Of course, the two sums may be different!

Elementary properties of series

Some properties of addition hold for series as well. For convergent series, given a real number c we have:

$$\sum_{s=0}^{+\infty} ca_s = c \sum_{s=0}^{+\infty} a_s \tag{6.4}$$

If $c \neq 0$, then (6.4) holds also for divergent series. In any case, *if* $c \neq 0$, *the series with general terms* a_n *and* $c \cdot a_n$ *have the same behaviour.*
Moreover, for convergent series, the following *additive property* holds:

$$\sum_{s=0}^{+\infty} (a_s + b_s) = \sum_{s=0}^{+\infty} a_s + \sum_{s=0}^{+\infty} b_s. \tag{6.5}$$

The property (6.5) may also be extended to divergent series if we exclude the case where the two sums on the right hand side diverge to a different infinity, like $(+\infty) + (-\infty)$. In all other cases, the property (6.5) has to be read in the following way: if one of the sums on the right hand side diverges and the other converges or diverges to the same infinity, then also the series on the left hand side diverges to the same infinity. We leave the easy verification of all these properties as an exercise.

6.2 Geometric series

A very important series, both from the theoretical and form the pratical point of view, is the geometric series with ratio q:

$$\sum_{k=0}^{+\infty} q^k = 1 + q + q^2 + \cdots + q^k + \cdots , \tag{6.6}$$

Let us first study the case $q = 1$. The terms of the series are all equal. The sequence of its partial sums is

$$S_n = \underbrace{1 + 1 + 1 + \cdots + 1}_{n+1 \text{ times}} = n + 1 \to +\infty$$

and therefore the series diverges.

If $q \neq 1$, by applying the formula for the sum of terms in a geometric progression, proven in Section 4.3 of Chapter 1, we have:

$$S_n = \sum_{k=0}^{n} q^k = 1 + q + q^2 + \cdots + q^n = \frac{1 - q^{n+1}}{1 - q}.$$

The behaviour of the series is then easily detected, because the asymptotic behaviour of S_n is determined by q^{n+1}. As seen in Section 1.6 of Chapter 3,

$$\lim_{n \to +\infty} q^{n+1} = \begin{cases} +\infty & \text{if } q > 1 \\ 0 & \text{if } -1 < q < 1 \\ \text{does not exist} & \text{if } q \leq -1 \end{cases},$$

from which we deduce the following conclusions:

$$\begin{cases} \text{if } q \geq 1 & \text{the series diverges to } +\infty \\ \text{if } -1 < q < 1 & \text{the series converges to the sum } \dfrac{1}{1-q} \\ \text{if } q \leq -1 & \text{the series is irregular.} \end{cases}$$

The series of example 1.1

$$\sum_{k=1}^{+\infty} \frac{1}{2^k}$$

is a geometric series with ratio $q = \dfrac{1}{2}$, and therefore converges to 1. Indeed, we may write [3]

$$\sum_{k=1}^{+\infty} \left(\frac{1}{2}\right)^k = \sum_{k=0}^{+\infty} \left(\frac{1}{2}\right)^k - \left(\frac{1}{2}\right)^0 = \frac{1}{1 - 1/2} - 1 = 1.$$

[3] Be careful with the initial index!

The geometric series

$$\sum_{s=0}^{+\infty}(-1)^s = 1-1+1-1+1-1+\cdots$$

with ratio $q=-1$ is irregular. Indeed, we have

$$S_n = \{1,0,1,0,1,\ldots\}\,.$$

The geometric series

$$\sum_{s=1}^{+\infty}(-2)^s = -2+4-8+16-\cdots\,,$$

with ratio $q=-2$ is irregular. Differently from the geometric series with ratio -1, here we have $\lim|S_n| = +\infty$.

If the first term of the series is not 1, the sum of the series (when $|q|<1$) may be computed by using the homogeneity property (6.4).

Therefore, for example, the geometric series

$$\sum_{s=2}^{+\infty}5e^{-3s} = \frac{5}{e^6}+\frac{5}{e^9}+\frac{5}{e^{12}}+\cdots\,,$$

with first term $5/e^6$ and ratio $1/e^3$ (which converges, because $1/e^3$ is between 0 and 1) has the sum (we collect the common factor $5/e^6$)

$$\frac{5e^{-6}}{1-e^{-3}} = \frac{5}{e^6-e^3} \simeq 0.013.$$

• (⇛ **Chapter 11**) *Fundamental value of a share.* To evaluate a share we may use the so-called Gordon's formula: this formula simply assigns to a share the sum of the discounted values of its future dividends. If this share pays d_1 after one year, d_2 after two years, d_3 after three years, ... d_s after s years, and so on, the value of the share is

$$\sum_{k=1}^{+\infty}\frac{d_k}{(1+r)^k},$$

where $(1+r)^k$ is the accumulation factor in compound interests, with annual interest rate r. Financial analysts usually assume that the dividends will increase regularly over time with growth rate g and that, therefore,

$$d_2 = d_1(1+g)\ ;\quad d_3 = d_2(1+g) = d_1(1+g)^2\ ;\ \ldots$$

Thus, in general, we have

$$d_k = d_1(1+g)^{k-1}\,.$$

In this situation, the value of the share is

$$\sum_{k=1}^{+\infty} \frac{d_1 (1+g)^{k-1}}{(1+r)^k}. \tag{6.7}$$

If we further assume that $0 < g < r$, so that $0 < \dfrac{1+g}{1+r} < 1$, the series converges and its sum is

$$\frac{d_1}{1+r} \frac{1}{1-\dfrac{1+g}{1+r}} = \frac{d_1}{r-g}. \tag{6.8}$$

This value represents the "*theoretical value of the share, at rate r*", as a function of the first dividend d_1 and of the expected growth rate g.

The very simple formula (6.8) states that the value of a share can be obtained as the ratio between the first dividend and the difference between the rate r (the usual market rate for investments) and the growth rate.

For example, a share with an expected first dividend 1000 and growth rate 1%, if evaluated at rate 11% has theoretical value

$$V = \frac{1000}{11\% - 1\%} = \frac{1000}{10\%} = 10000.$$

Notice that constant dividends $d_1 = d_2 = \cdots = d_s = \cdots = d$ (that is, $g = 0$), imply

$$V = \frac{d}{r},$$

which is the formula currently used to evaluate pieces of land, which have a constant "dividend" d over time. It gives a first estimate of the value of a piece of land, as the ratio between the annual income d and the interest rate r.

In order to deduce Gordon's formula, we assumed $g < r$. Let us now see what will happen if, forgetting to verify this assumption, we were still to use this formula with $g > r$. If, as above, $d_1 = 1000$, $r = 11\%$ and $g = 12\%$ (much better prospects of growth), Gordon's formula $V = d_1/(r-g)$ will give, instead of a value greater than 10000 (as the better prospects suggest), a negative value

$$P = \frac{1000}{11\% - 12\%} = -10000,$$

which is completely absurd! Actually, if $g > r$, then $\dfrac{1+g}{1+r} > 1$ and the series *diverges*.

6.3 The problem of convergence

The example illustrated in the previous pages may mislead the reader. Indeed, one may think that it is possible to decide the convergence of a series in a natural way by following the definition, and therefore by computing

$$\lim_{n \to +\infty} S_n.$$

This way we can also compute the sum of the series, which is the value of the limit if this limit exists.

Unfortunately, this way is usually precluded, because only in a very few cases we can obtain a handy expression of the partial sum S_n. To decide if a series converges or not is therefore often an indirect problem. This means that we have to search for some structural properties of $\{a_n\}$, the sequence of the terms of the series, which guarantee that the series converges (or not).

In the case of convergence, in order to compute the sum S in an application, we may use an approximation of S, usually given by a partial sum S_n, with n large enough (and a calculator...).

6.3.1 A necessary condition for convergence

We start with a simple necessary condition for convergence. This condition does not guarantee the convergence of the series, but, if not satisfied, it excludes it.

Theorem 3.1. *If* $\sum_{n=0}^{+\infty} a_n$ *converges, then the general term tends to* 0:

$$\lim_{n\to+\infty} a_n = 0.$$

Proof. It is enough to observe that, by definition (6.3), we have

$$a_n = S_n - S_{n-1}.$$

By passing to the limit for $n \to \infty$, since both S_n and S_{n-1} converge to S (the sum of the series), their difference (which is a_n) must converge to $S - S = 0$, and therefore be infinitesimal:

$$\lim_{n\to+\infty} a_n = \lim_{n\to+\infty} S_n - \lim_{n\to+\infty} S_{n-1} = S - S = 0. \square$$

The Mengoli series converges; it is therefore not surprising that its general term $1/n\,(n+1)$ is infinitesimal. The geometric series converges only when its general term is infinitesimal. The fact that the general term tends to zero *does not* ensure, in general, that the series converges. For example, for the so-called *harmonic series*

$$\sum_{k=1}^{+\infty} \frac{1}{k} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots + \frac{1}{k} + \cdots, \tag{6.9}$$

the general term[4] $a_n = 1/n$ is infinitesimal, but nevertheless one can prove that $S_n \sim \ln n \to +\infty$.

[4] The name *harmonic* is due to the fact that the general term $1/n$ is the harmonic mean between the preceding term $1/(n-1)$ and the following one $1/(n+1)$. We recall that the *harmonic mean* between two numbers a and b is

$$\frac{2}{1/a + 1/b} = \frac{2ab}{a+b}.$$

6.4 Series with non-negative terms

In this section we describe some methods for determining the behaviour of series having terms with constant sign, in particular positive. If we further recall that, altering a finite number of terms, the character of a series does not change, we may conclude that the properties which hold for series with positive terms hold also for series with terms having a *constant sign* for all terms a_n corresponding to values of n which are large enough. Let us therefore consider series

$$\sum_{n=0}^{+\infty} a_n$$

where all terms a_n are positive or at least non negative ($a_n \geq 0$). A first result excludes that these series may be irregular.

Proposition 4.1. *Every series with non negative terms is regular.*

Proof. From equation (6.3) we deduce that, if $a_n \geq 0$, for every n

$$S_n = S_{n-1} + a_n \geq S_{n-1}, \quad \text{for every } n.$$

The sequence S_n is increasing and therefore, using the regularity theorem for a monotonic sequence, it either converges or diverges to $+\infty$. □

We now introduce the so-called comparison criteria, which are very useful in many circumstances. Let us consider the series

$$\sum_{n=0}^{+\infty} a_n \quad \text{and} \quad \sum_{n=0}^{+\infty} b_n$$

with $a_n \geq 0$, $b_n \geq 0$.

Theorem 4.2 (Comparison criterion). *If*

$$0 \leq a_n \leq b_n, \quad \textit{for every} \quad n, \tag{6.10}$$

then

$$\sum_{n=0}^{+\infty} b_n < +\infty \textit{ implies } \sum_{n=0}^{+\infty} a_n < +\infty \quad \textit{and} \quad \sum_{n=0}^{+\infty} a_n = +\infty \textit{ implies } \sum_{n=0}^{+\infty} b_n = +\infty$$

If equation (6.10) holds, the first series is called a *minorant* of the second one and the second series is called a *majorant* of the first one.

Proof. Let A_n and B_n be the partial sums of the two series. Equation (6.10) implies that

$$0 \leq A_n \leq B_n, \qquad \text{for every } n.$$

If the majorant series converges, then also the minorant one converges. Indeed, if B_n converges, then A_n also converges, by applying the comparison theorem for limits

of sequences (section 5.3 of Chapter 3). In an analogous way, if the minorant series diverges, then also the majorant one diverges. □

Often, the following corollary is much more useful and convenient:

Corollary 4.3 (Asymptotic comparison criterion). *If*

$$a_n \sim b_n,$$

then $\sum_{n=0}^{+\infty} a_n$ *and* $\sum_{n=0}^{+\infty} b_n$ *have the same behaviour.*

Proof. The relation $a_n \sim b_n$ means that for $n \to +\infty$

$$\frac{a_n}{b_n} \to 1.$$

Therefore, for every $n \geq N$ (N large enough), we have

$$\frac{1}{2} < \frac{a_n}{b_n} < 2,$$

and consequently

$$\frac{1}{2} b_n < a_n < 2b_n. \tag{6.11}$$

Equation (6.11) proves that $\sum_{n=0}^{+\infty} a_n$ is both a minorant and a majorant of the series with general terms $b_n/2$ and $2b_n$, which have the same behaviour as $\sum_{n=0}^{+\infty} b_n$. □

Example 4.1. Series of particular importance are

$$\sum_{k=1}^{+\infty} \frac{1}{k^\alpha}, \tag{6.12}$$

where α is a real and positive parameter. These series are called *generalized harmonic series*. For $\alpha = 1$ the series (6.12) is the harmonic series (6.9). One may prove[5] that

$$\sum_{k=1}^{+\infty} \frac{1}{k^\alpha} \quad \begin{cases} \text{converges} & \text{if } \alpha > 1 \\ \text{diverges} & \text{if } \alpha \leq 1. \end{cases}$$

For example, $\sum_{k=1}^{+\infty} \frac{1}{\sqrt{k^3}}$ converges and $\sum_{k=1}^{+\infty} \frac{1}{\sqrt{k}}$ diverges. We now limit ourselves to prove the case $\alpha \geq 2$. If $\alpha = 2$, since

$$\frac{1}{k^2} \sim \frac{1}{(k-1)k},$$

[5] This criterion describes the behaviour of generalized harmonic series and can be proved as a direct application of the *integral comparison criterion*, stated in section 8 of Chapter 7.

by applying corollary 4.3, the series $\sum_{k=1}^{+\infty} \frac{1}{k^2}$ has the same behaviour as the Mengoli series[6]. Therefore it converges. If $\alpha > 2$, we have $n^2 \leq n^\alpha$ and therefore

$$\frac{1}{n^\alpha} \leq \frac{1}{n^2}.$$

We have just seen that the series $\sum_{k=1}^{+\infty} \frac{1}{k^2}$ converges. Therefore, by theorem 4.2, all series with $\alpha > 2$ also converge, since they are minorants of this one.

Example 4.2. Consider the series

$$\sum_{n=1}^{+\infty} \frac{n + 5\ln n + 18}{n^2 + 14e^{-n}}. \tag{6.13}$$

Its general term is quite complicated. Since the first term in both the numerator and the denominator are the predominant ones, we have

$$\frac{n + 5\ln n + 18}{n^2 + 14e^{-n}} \sim \frac{n}{n^2} = \frac{1}{n},$$

which is the general term of the harmonic series. The harmonic series diverges and therefore, by corollary 4.3, series (6.13) also diverges.

Criterion of the ratio

We know that for the terms of a harmonic series the ratio between the $(k+1)$-th term and the k-th term is constant:

$$\frac{a_{k+1}}{a_k} = \frac{q^{k+1}}{q^k} = q. \tag{6.14}$$

We suppose that the ratio q is positive and that $q < 1$, so that the geometric series has positive terms and converges. Equation (6.14) says that, by passing from the k-th term to the $(k+1)$-th term, we have a contraction with ratio (percentually) equal to q. Suppose that, for a series $\sum_{n=0}^{+\infty} a_k$ with positive terms, a relation

$$\frac{a_{k+1}}{a_k} \leq q \tag{6.15}$$

with $q < 1$ holds for all k large enough. This relation tells us that the terms contract at least as quickly as the terms of a convergent geometric series. By applying the comparison criterion, this series also converges.

Analogously, if for a series with positive terms $\sum_{n=0}^{+\infty} a_k$ a relation of type

$$\frac{a_{k+1}}{a_k} \geq q \tag{6.16}$$

[6] See example 1.2.

with $q > 1$ holds for k large enough, by the comparison criterion this series also diverges. In this case, we also see that the general term of the series cannot be infinitesimal and therefore the non convergence of the series may be deduced by the fact that the general necessary condition for convergence is not satisfied.

How can we easily show that condition (6.15) or condition (6.16) holds? An attempt which sometimes has success is suggested by the following

Theorem 4.4 (Criterion of the ratio for series). *Let*

$$\lim_{k\to+\infty} \frac{a_{k+1}}{a_k} = L.$$

If $L < 1$ *the series converges, if* $L > 1$ *the series diverges.*

Warning! If $L = 1$ we cannot deduce anything about convergence in general. For example, the generalized harmonic series has the ratio between consecutive terms equal to

$$\frac{1/(n+1)^\alpha}{1/n^\alpha} = \left(\frac{n+1}{n}\right)^{-\alpha} = \left(1+\frac{1}{n}\right)^{-\alpha} \to 1, \quad \text{for every } \alpha$$

and the series has different behaviours according to the value of α (it converges for $\alpha > 1$, it diverges otherwise).

Example 4.3 (*Exponential series*). We consider the series

$$\sum_{k=0}^{+\infty} \frac{x^k}{k!} \tag{6.17}$$

which depends on the real parameter x. For $x = 0$ the series clearly converges. For $x > 0$ the series has positive terms. The ratio between two consecutive terms is

$$\frac{x^{k+1}/(k+1)!}{x^k/k!} = \frac{x}{k+1} \to 0 \quad \text{for every } x.$$

Thus the series converges, because the limit of this ratio is less than 1. This series is called the *exponential series* because its sum is the exponential function.

$$\boxed{\sum_{k=0}^{+\infty} \frac{x^k}{k!} = e^x}$$

• *The law of small numbers.* We consider the exponential series (6.17) with $x > 0$ and we "normalize" it by dividing every term by its sum e^x. We obtain:

$$\sum_{k=0}^{+\infty} \frac{x^k}{k!} e^{-x} = 1 \tag{6.18}$$

In the insurance field the terms of this series are exceptionally important because they represent the probability that in a certain period of time some events, often called "rare events", happen.

For example, let us consider an insurance policy for cars. We ask how many car accidents, with liability for damages, an insured driver may cause in a year. This number cannot be predicted: one can only say that it may be 0 or 1 or 2 or 3 and so on. We say that the number of car accidents is a *random number*. It is quite a crucial matter for the insurance company to answer the question: what probability must be assigned to each of these possibilities? This means studying the so-called *probability distribution of the random number* of car accidents.

The answer is: if x is the mean number of accidents caused by the insured driver, the probability that the number of accidents is $0, 1, 2, \ldots$ is given by the so-called "law of small numbers", also called Poisson's[7] "law of rare events".

The probability that the number of accidents is exactly k is the term of index k in the series (6.18), that is[8]

$$\Pr\big(\text{number of accidents} = k\big) = \frac{x^k}{k!}e^{-x}.$$

In practice, we first have to estimate x. If the insured driver caused 2 accidents in 30 years of driving, we may estimate

$$x = \frac{2}{30} = \frac{1}{15} = 6.\overline{6}\%.$$

We may then compute the different probabilities in which we are interested:

$$\begin{aligned}
\Pr\left(\text{number of accidents} = 0\right) &= \frac{(1/15)^0}{0!}e^{-1/15} \simeq 93.56\% \\
\Pr\left(\text{number of accidents} = 1\right) &= \frac{(1/15)^1}{1!}e^{-1/15} \simeq 6.24\% \\
\Pr\left(\text{number of accidents} = 2\right) &= \frac{(1/15)^2}{2!}e^{-1/15} \simeq 0.2\%
\end{aligned}$$

and so on. The condition (6.18) simply guarantees that the sum of all the probabilities assigned to the various number of car accidents is 1, as it should be.

6.5 Series with terms of non-constant sign

At least in principle, the fact that the terms of a series may have any sign produces some complications. In many cases we may overcome them with a criterion, which reduces the study of a series with non-constant sign to that of a series with non-negative terms.

Theorem 5.1 (Absolute convergence criterion). *If the series* $\displaystyle\sum_{k=0}^{+\infty} |a_k|$ *converges, then the series* $\displaystyle\sum_{k=0}^{+\infty} a_k$ *also converges.*

[7] Siméon Denis Poisson (1781-1840).

[8] Pr means *probability.*

If the series $\sum_{n=0}^{+\infty} |a_k|$ converges, we say that the series $\sum_{n=0}^{+\infty} a_k$ *converges absolutely*. Therefore we may rephrase the above conclusion as follows: *if a series converges absolutely, then it converges.*

The converse does not hold: *a series may converge but not absolutely.*

For instance, the series

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \cdots = \sum_{s=0}^{+\infty} \frac{(-1)^s}{s+1}$$

converges, as we shall see later on, but the series of the absolute values of its terms diverges (it is the harmonic series). The convergence of the series is due, essentially, to the compensation between positive and negative terms.

Example 5.1. Let us reconsider the exponential series. For $x < 0$, its terms alternately have positive and negative signs. On the other hand, the series of the absolute values

$$\sum_{k=0}^{+\infty} \frac{|x|^k}{k!}$$

corresponds to an exponential series with positive terms, which we know to converge. Thus the series converges absolutely and therefore it converges.

6.5.1 Series with terms of alternate sign

This expression denotes series whose terms are alternately positive and negative (at least starting from a certain term). For example, the series

$$\sum_{n=0}^{+\infty} \frac{(-1)^n}{n+1} \tag{6.19}$$

is a series with *terms of alternate sign*. It is possible to spot some characteristics of such series which guarantee their convergence. Let us denote the absolute values of the terms of a series with alternate signs by $u_n > 0$. If the series starts with a positive term, we may rewrite it as:

$$u_0 - u_1 + u_2 - u_3 + \cdots + (-1)^n u_n + \cdots = \sum_{n=0}^{+\infty} (-1)^n u_n.$$

The following theorem (known as Leibniz's criterion) holds.

Theorem 5.2. *Consider the series*

$$\sum_{n=0}^{+\infty} (-1)^n u_n, \quad \textit{with} \quad u_n > 0.$$

If the following conditions hold:

(i) $u_{n+1} \leq u_n$ *for every* n,
(ii) $\lim_{n \to +\infty} u_n = 0$,
then the series converges.

Properties (i) and (ii) are satisfied by the series (6.19), which therefore converges[9].

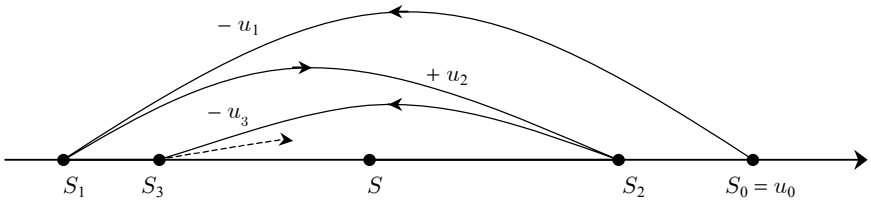


Figure 6.1.

Let us try to understand why the two conditions (i) and (ii) guarantee convergence. The first partial sum is $S_0 = u_0 > 0$. The second one is $S_1 = S_0 - u_1 \leq S_0$, but still non-negative, since $u_0 \geq u_1$. The third partial sum is $S_2 = S_1 + u_2 \geq S_1$. Since $u_2 \leq u_1$, the third partial sum is still less than or equal to S_0. Proceeding this way, we note that the even partial sums $S_0, S_2, S_4, \ldots$ form a decreasing sequence, while the odd partial sums $S_1, S_3, S_5, \ldots$ form a decreasing sequence:

$$\begin{aligned} S_0 &\geq S_2 \geq S_4 \geq \ldots \\ S_1 &\leq S_3 \leq S_5 \leq \ldots \end{aligned}$$

Moreover, the terms of the first sequence (the even partial sums) are all greater or equal to the terms of the second sequence (the odd partial sums) and the interval between S_{2n} and S_{2n-1} (that is $S_{2n} - S_{2n-1} = u_{2n}$) has infinitesimal length because of property (ii). This forces $\{S_n\}$ to converge to a real number S for $n \to +\infty$.

6.6 Exercises

6.1. *Choose the right answer.* To determine the behaviour of a series $\sum_{n=0}^{+\infty} a_n$ means to check that: (a) the general term a_n tends to zero; (b) the series converges, diverges or is irregular; (c) the series gets on well with everybody.

6.2. *Choose the right answer.* A series $\sum_{n=0}^{+\infty} a_n$ converges and has sum S if: (a) its general term a_n tends to zero; (b) its general term a_n tends to S; (c) the sequence of its partial sums $S_n = \sum_{n=0}^{+\infty} a_k$ tends to S.

[9] For the interested reader: with methods which are less elementary than the one presented here, one can prove that the series has sum $\ln 2$.

6.3. Study the behaviour of the series

$$\sum_{n=1}^{+\infty} \frac{n+1}{n^2+1}, \quad \sum_{n=2}^{+\infty} \frac{n+1}{n^3-1}, \quad \sum_{n=0}^{+\infty} n^{100} e^{-n}, \quad \sum_{n=2}^{+\infty} (-1)^n \frac{1}{\ln n}.$$

6.4. Study the behaviour of the series

$$\sum_{n=1}^{+\infty} \frac{n+5e^{-n}}{2n^2+3\ln n}, \quad \sum_{n=1}^{+\infty} \frac{\ln n}{n\sqrt{n}}, \quad \sum_{n=1}^{+\infty} \frac{\sin n}{n^2}, \quad \sum_{n=1}^{+\infty} \frac{(-1)^n}{n^2+9\sin n}.$$

6.5. Determine the values of x for which the following series converges:

$$\sum_{n=0}^{+\infty} \frac{x^n}{n+1}.$$

6.6. A particle moves from point $A_0 = (1, 0)$ along the trajectory A_0, A_1, $A_2, \ldots$ $A_n, \ldots$, as shown in the following picture (the segments A_0A_1, $A_2A_3 \ldots$ are perpendicular to the bisector). What is the length of the polyline $A_0A_1A_2A_3$? What is the length of the whole polyline?

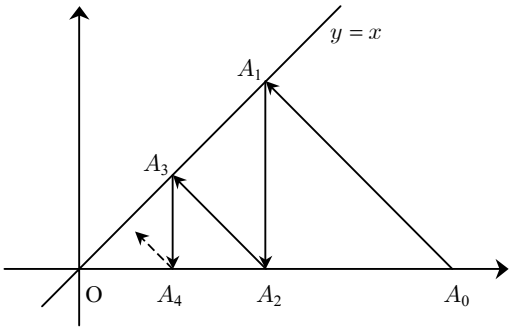


Figure 6.2.

6.7. Consider a *perpetuity*: i.e., the amounts R_n of money spread in time from now up to... eternity. The present value of a perpetuity is the sum of the present values at compound interest rate i of the annual incomes $R_1, R_2, \ldots, R_n, \ldots$ at the years $1, 2, \ldots, n, \ldots$ and, therefore, it is given by the series

$$V = \sum_{n=1}^{+\infty} \frac{R_n}{(1+i)^n}.$$

Calculate V in the following cases:

— $R_n = R$ constant $= 20$, $i = 5\%$;

— $R_n = R_1 (1+g)^{n-1}$, variable incomes in a geometric progression of ratio $1+g$, with $R_1 = 20$, $i = 5\%$, $g = 2\%$.

7

Integral Calculus

Many issues in Probability, Statistics and Finance lead in a natural way to *integral calculus* and emphasize the need to deal with this topic.

The concept of *integral* is the third one we meet in the so-called *infinitesimal calculus*, after the concepts of *limit* and *derivative*. Historically, however, these three notions were introduced in exactly the reverse order.

The notion of limit was introduced in a rough form by D'Alembert (18th century) and later developed by Cauchy and Weierstrass (19th century). Differential calculus, which is linked with many problems of a physical and geometrical nature, began in the 17th century with Fermat and Descartes and was then developed by Newton and Leibniz.

The notion of integral is linked with important geometric problems, such as the "squaring of the circle" (squaring stands here for the possibility of computing a formula for the area) and may be dated back to Archimedes[1] and his "method of exhaustions".

The chapter develops along the following lines.

• We give an introductory example, dealing with the computation of an area, and this leads us to the definition of the *Riemann integral.*

• We then consider one of the most important issues of the whole integral calculus: its connection with the differential calculus. The core result is expressed by the *first fundamental theorem*, which also provides an explicit formula for computing the integral of a function.

[1] Archimedes of Siracusa, 287–212 B.C.

• We introduce the notion of an "indefinite integral" and we describe the main methods for its computation.

• The definition of "Riemann integral" is then extended to unbounded functions or to functions defined over unbounded intervals: the so-called "improper Riemann integral". This notion is particularly important for applications in Probability or Statistics.

• The last part is devoted to the study of integral functions, i.e. integrals of functions defined over a variable interval. The main result is the *second fundamental theorem*, which describes the relation between a derivative and an integral in this more general context.

7.1 Introduction

To illustrate the concept of integral, we start from the problem of computing the area of a plane figure.

Let $f : [a, b] \to \mathbb{R}$ be a positive function, with a graph as in Figure 1: we want to compute the area of the plane region bounded by the graph of f, the x-coordinate axis, and the lines $x = a$, $x = b$. Such a figure is called a *trapezoid*. In order to compute its area, the idea is to approximate such a figure in the following way: we split it into vertical slices (trapezoids), which are so thin that the function f is nearly constant along the x-values occurring in each slice. We then take an x-value for each slice, and we approximate this slice with a rectangle of height equal to the value that f takes in the chosen x-value. Finally we compute the area of each rectangle and we add all these areas.

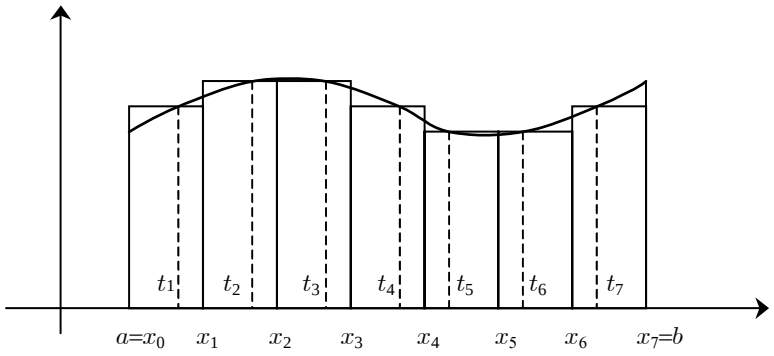


Figure 7.1. Approximation of the trapezoid area

To be precise, we subdivide the interval $[a, b]$ into n intervals $[x_{s-1}, x_s]$ of equal length $\Delta x = (b - a)/n$, with $s = 1, 2, ..., n$ and $x_0 = a$, $x_n = b$. We then choose a point t_s in each interval $[x_{s-1}, x_s]$ and we consider the rectangle of base $[x_{s-1}, x_s]$ and height $k_s = f(t_s)$.

We compute the area $k_s \Delta x$ of each rectangle and we sum up these areas:

$$\sum_{s=1}^{n} f(t_s)(x_s - x_{s-1}) = \sum_{s=1}^{n} k_s \Delta x \qquad \left(\Delta x = x_s - x_{s-1} = \frac{b-a}{n}\right). \tag{7.1}$$

If the intervals $[x_{s-1}, x_s]$ are small enough and f does not vary too much in each interval (see below), the sum (7.1) is an approximation of the value of the area we intend to compute. At this point, we hope that, by refining and refining the subdivision of the interval $[a, b]$, these approximations tend to the actual value of the trapezoid area.

When is this true? Our hope is based on the fact that our choice of the points t_s in $[x_{s-1}, x_s]$ *must be irrelevant*. Otherwise, such a limit would be completely arbitrary and, therefore, of no interest at all. Such a behaviour is strictly connected with the behaviour of the function f: everything works well if f is nearly constant on every interval of very small length, which is exactly the property of *continuity* of f! We can actually allow that f be discontinuous at some isolated points (as in a jump discontinuity), but we have to require that f be bounded along all the intervals $[x_{s-1}, x_s]$. If this is the case, we may then expect that, by indefinitely subdividing the interval $[a, b]$, the value of the approximating sums (7.1) tends to a well defined value S. This value S is called the *Riemann*[2] *definite integral* of the function f on $[a, b]$.

7.2 The Riemann integral

Let us formalize our above reasoning in order to obtain the definition of a *Riemann integral*. Let f be a function, defined over an interval $[a, b]$ and *bounded* on that interval. We subdivide the interval $[a, b]$ into n smaller intervals $[x_{i-1}, x_i]$, $i = 1, \ldots, n$, having equal width $\Delta x = (b-a)/n$. In each of these intervals we then choose a point $t_i \in [x_{i-1}, x_i]$ and we consider the sum, called the *Riemann integral sum*,

$$\sigma_n = \sum_{i=1}^{n} f(t_i)\Delta x = \frac{b-a}{n} \sum_{i=1}^{n} f(t_i).$$

We stress again that the sum σ_n depends not only on n (the number of subintervals in the subdivision), but also on the choice of the points t_i, $i = 1, \ldots, n$.

At this point we may increase the value n, obtaining subintervals of smaller and smaller width: we let n go to $+\infty$. We are ready to define the notion of integral.

Definition 2.1 (Riemann integral). *The* ***integral*** *of f in $[a, b]$ is the limit, for $n \to +\infty$, of the integral sum $\sigma_n = \dfrac{b-a}{n} \displaystyle\sum_{i=1}^{n} f(t_i)$, provided that such a limit exists*

[2] Georg Friederich Bernard Riemann (1826-1866), German mathematician.

and does not depend [3] *on the choice of the points* t_i. We write:

$$\boxed{\int_a^b f(x)\mathrm{d}x := \lim_{n\to+\infty} \sigma_n = \lim_{n\to+\infty} \frac{b-a}{n} \sum_{i=1}^{n} f(t_i).} \tag{7.2}$$

Two questions arise spontaneously:

a) Consider all the functions defined and bounded on $[a, b]$: are they all "integrable", i.e., does there always exist a limit as in the above definition which does not depend on the choice of the points t_i? The disenchanted reader should have already understood that this is not true: there are functions for which it is possible to speak about their integral, which we call *integrable functions*, and others which behave in an irregular way: either the limit does not exist or it depends on the choice of the points t_i. But then: which are the *integrable functions*?

b) Suppose now that f is integrable in $[a, b]$. How can we compute its integral? By using the above definition 2.1?

In this introduction to the integral calculus we have to limit ourselves to some partial answers to the above questions.

Some classes of integrable functions

Recall the introductory example. The *integrability* of a function f in $[a, b]$ depends essentially on the property that f is *bounded* and "not too discontinuous". The boundedness of f alone is not enough: indeed there exist some "pathological" functions, which are bounded but have too many discontinuity points, and are therefore not integrable. Let us examine some classes of functions which are certainly integrable.

• The functions which are *continuous in* $[a, b]$ are integrable. Note that, in this case, such a function f is surely *bounded*, since the interval $[a, b]$ is closed.

• Functions which are *bounded* and *discontinuous only at a finite set of points* are integrable. In particular, piecewise constant functions are integrable.

A function with a step graph as in figure 2 satisfies

$$\int_a^b f\,(x)\,\mathrm{d}x = k_1(x_1 - x_0) + k_2(x_2 - x_1) + k_3\,(x_3 - x_2) + k_4\,(x_4 - x_3)\,.$$

• Functions which are *monotonic in* $[a, b]$ are integrable. Note that such functions may also have a countable infinity of discontinuity points.

Therefore, all of the elementary functions we met in our calculus course are integrable, such as, for example, the following: polynomials, exponential functions,

[3] To be precise, we should clearly explain the meaning of the condition that *the limit is independent of the choice of the points* t_i, but we prefer to gloss over this question, of a rather technical nature.

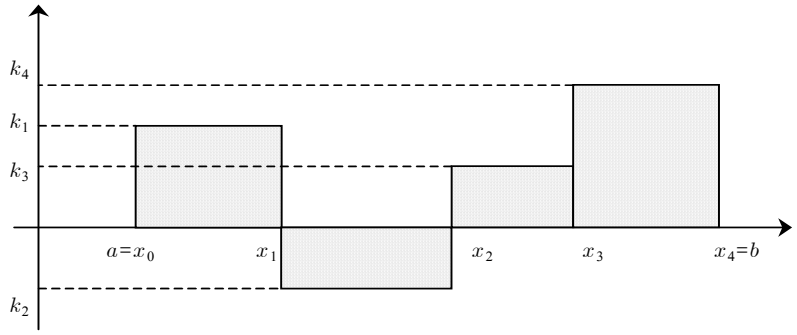


Figure 7.2.

trigonometric functions (in every interval $[a, b] \subset \mathbb{R}$ on which they are defined), and logarithmic functions (in every interval $[a, b] \subset (0, +\infty)$).

We conclude this section with some useful remarks.

Let be f integrable in $[a, b]$. Changing the value of f at a finite numbers of points, the integral of f does not change.

Indeed, if we consider the graph of a function f and we remove a point, say $(\alpha, f(\alpha))$, the area of the part of the plane which lies under the graph will not change. We can also remove more than one point, even a countable infinity of points, and the area will remain the same. The same happens if, once we have removed a point $(\alpha, f(\alpha))$ from the graph, we replace it with another point (α, y_α) of different height, such as, for example, $(\alpha, f(\alpha) + 10)$.
In particular, for a bounded function, it is equivalent to speak of its integral on a closed interval $[a, b]$, or on an open one (a, b), or on any of the intervals $[a, b)$, $(a, b]$.

We note however that a continuous function over a non-closed interval is not necessarily bounded: in this case therefore the mere continuity of the function is not sufficient, and we need to verify that f is actually bounded.

• *Notation of integral*. The mathematical symbol of integral

$$\int_a^b f(x)\mathrm{d}x \tag{7.3}$$

recalls how the integral is defined:

The symbol "$\int_a^b$" is a stretched out "S", which recalls the summation symbol $\sum_{i=1}^n$, while the product $f(x)\mathrm{d}x$ evokes the generic addend $f(t_i)\Delta x$ in this summation:

$$\begin{array}{ccc} \displaystyle\sum_{i=1}^{n} & f(t_i) & \Delta x \\ \updownarrow & \updownarrow & \updownarrow \\ \displaystyle\int_a^b & f(x) & \mathrm{d}x \end{array}$$

The choice of the name "x" for the *integration variable* is completely arbitrary. As an example, if we change this name to t we get the equivalent form

$$\int_a^b f(t)\,\mathrm{d}t.$$

For this reason, the integration variable is said to be a *dummy variable*: by substituting it with another one nothing changes, neither formally nor substantially. So to say, this variable plays the same role as the index in a summation. Thus, the integral of f in $[a,b]$ could be simply denoted by

$$\int_a^b f.$$

7.3 Properties of the integral

7.3.1 Additivity, linearity, monotonicity

We now state some properties of the integral, which have a clear geometrical meaning.

• *Additivity with respect to the integration interval.* We consider two adjacent intervals $[a,b]$ and $[b,c]$ and a function f which is integrable in $[a,c]$. We have

$$\int_a^b f(x)\,\mathrm{d}x + \int_b^c f(x)\,\mathrm{d}x = \int_a^c f(x)\,\mathrm{d}x \tag{7.4}$$

The geometrical interpretation of the integral makes the assertion obvious. It is a common convention to define, for $b < a$, the following integral:

$$\int_a^b f(x)\,\mathrm{d}x := -\int_b^a f(x)\,dx$$

We note that, in this way, the integral depends on the x-coordinate axis orientation. As an immediate consequence, we have

$$\int_a^a f(x)\,\mathrm{d}x = 0,$$

which was intuitively clear. Another consequence is that the formula (7.4) also holds when b does not lie between a and c. The property expressed by (7.4) is called *additivity w.r.t. the integration interval*.

• *Linearity.* If f_1, f_2 are integrable in $[a,b]$ and c_1, c_2 are two constants, then

$$\int_a^b [c_1 f_1(x) + c_2 f_2(x)]\,\mathrm{d}x = c_1 \int_a^b f_1(x)\,\mathrm{d}x + c_2 \int_a^b f_2(x)\,\mathrm{d}x.$$

In words: a linear combination of integrable functions is again integrable and its integral is the linear combination of the integrals. In particular

$$\int_a^b cf(x)\,\mathrm{d}x = c\int_a^b f(x)\,\mathrm{d}x.$$

• *Positivity.* Suppose that f is integrable and non-negative in $[a,b]$:

$$f(x) \geq 0 \text{ for all } x \in [a,b].$$

Then:

$$\int_a^b f(x)\,\mathrm{d}x \geq 0.$$

•*Monotonicity.* More generally, given two integrable functions in $[a,b]$ with

$$f_1(x) \leq f_2(x) \text{ for all } x \in [a,b],$$

then

$$\int_a^b f_1(x)\,\mathrm{d}x \leq \int_a^b f_2(x)\,\mathrm{d}x.$$

In particular, if f is integrable then also $|f|$ is integrable and

$$\left|\int_a^b f(x)\,\mathrm{d}x\right| \leq \int_a^b |f(x)|\,\mathrm{d}x.$$

If $b \leq a$, the property of linearity still holds, while the properties of positivity and monotonicity are modified in an evident way.

7.3.2 Mean value theorem

Let f be an integrable function in $[a,b]$. Another important property of the integral regards the so-called *mean value* of f in $[a,b]$, defined by the formula

$$M_f := \frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x.$$

The number M_f is a generalisation of the ordinary arithmetic mean. Indeed, using the definition of integral, we have

$$\frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x = \frac{1}{b-a}\lim_{n\to+\infty}\frac{b-a}{n}\sum_{i=1}^{n} f(t_i) = \lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n} f(t_i),$$

and the sum $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} f(t_i)$ is nothing but the arithmetic mean of the n numbers $f(t_1)$, $f(t_2), ..., f(t_n)$. The following theorem asserts that, given a continuous f, the mean value M_f always lies between the minimum and the maximum of f and that it has to be a value taken by the function f.

Theorem 3.1. *If f is continuous in $[a,b]$, there exists at least one point $c \in [a,b]$ such that*

$$f(c) = \frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x. \tag{7.5}$$

Equation (7.5) admits the following equivalent form

$$\int_a^b f(x)\,\mathrm{d}x = f(c)(b-a)$$

and, if f is positive, it has the following geometrical interpretation: by an appropriate choice of $c \in [a,b]$, the trapezoid determined by the graph of f is equivalent to a rectangle with base $(b-a)$ and height $f(c)$.

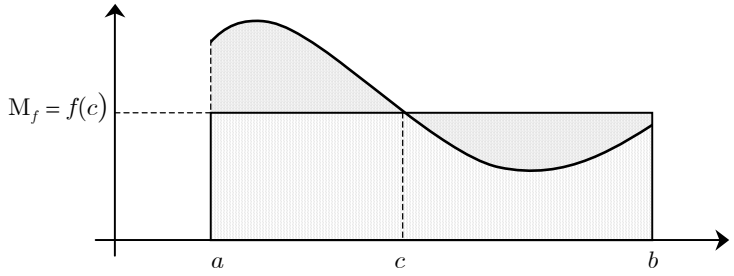


Figure 7.3. Geometrical meaning of the mean value theorem

Proof. We observe that:

$$\min_{x\in[a,b]} f(x) \le f(x) \le \max_{x\in[a,b]} f(x), \quad \text{for all } x \in [a,b].$$

By integrating on $[a,b]$ and using the monotonicity property, we obtain

$$\min_{x\in[a,b]} f(x)\cdot(b-a) \le \int_a^b f(x)\,\mathrm{d}x \le \max_{x\in[a,b]} f(x)\cdot(b-a),$$

and therefore

$$\min_{x\in[a,b]} f(x) \le \frac{1}{b-a}\int_a^b f(x)\,\mathrm{d}x \le \max_{x\in[a,b]} f(x). \tag{7.6}$$

Equation (7.5) is now a direct consequence of (7.6) and of the property of intermediate values for continuous functions (Darboux's Theorem): indeed (7.6) states that the mean value M_f lies between the minimum and the maximum of f, while the above mentioned property guarantees that such a value is one of those taken by f while x varies in $[a,b]$, that is, that there exists at least a point $c \in [a,b]$ such that $f(c) = M_f$. □

We note that the mean value of integral calculus we have introduced above has nothing to do with the Lagrange mean value we saw in Chapter 5.

7.4 The Fundamental Theorem of Calculus

Let us now consider the problem of the actual computation of an integral. It is practically impossible to obtain workable expressions for the integral sums σ_n, in order

to compute the limit of their sequence. For this reason, the definition of Riemann integral is never used to actually compute an integral.

The computational method, which is in most cases the best one, avoids the direct use of the definition and is based on a profound relation between integrals and derivatives. The notion of *antiderivative* will be of crucial importance.

Definition 4.1. *If G is a differentiable function in an interval (a, b) and*

$$G'(x) = f(x) \quad \text{for all } x \in (a, b)$$

the function G is called an ***antiderivative*** *of f in (a, b).*

For example, $G(x) = x^2$ is an antiderivative of $f(x) = 2x$, while $G(x) = \cos x$ is an antiderivative of $f(x) = -\sin x$.

Since the derivative of a constant is zero, two functions which differ by a constant have the same derivative. We deduce that, if a function has an antiderivative, then it has infinitely many.

For example, the function $f(x) = 1$ has as antiderivatives the infinite family of functions $G_k(x) = x + k$.

We shall see in the next section that, *given an antiderivative G of f in an interval (a, b) all the other antiderivatives of f are of the form $G + c$, $c \in \mathbb{R}$.*

We are now ready to introduce the *Fundamental Theorem of Calculus.*

Theorem 4.1. *Let G be an antiderivative of f in $[a, b]$. If f is integrable in $[a, b]$, then*

$$\boxed{\int_a^b f(x)\mathrm{d}x = G(b) - G(a).} \tag{7.7}$$

In words: the integral of f in $[a, b]$ is equal to the variation of G between a and b, usually denoted by the expression $[G(x)]_a^b$. Formula (7.7) is called the *formula for the calculation of an integral through the variation of an antiderivative.* It indeed reduces the difficult task of calculating an integral to the problem of finding an antiderivative. This theorem is therefore of fundamental importance for the integral calculus.

Proof. We consider a partition of $[a, b]$, i.e. a subdivision into n adjacent intervals of equal width $(b - a)/n$. Such a partition is determined by the points

$$x_0 = a, \quad x_1 = a + \frac{b-a}{n}, \ldots, \quad x_{n-1} = a + (n-1)\frac{b-a}{n}, \quad x_n = b.$$

We subtract and add $G(x_i)$ for $i = 1, \ldots, n-1$ to $G(b) - G(a)$ and we obtain

$$G(b) - G(a) =$$

$$= G(b) - G(x_{n-1}) + G(x_{n-1}) + \cdots - G(x_1) + G(x_1) - G(a) =$$

$$= \sum_{i=1}^{n} \left[G(x_i) - G(x_{i-1})\right].$$

Lagrange's mean value theorem allows us to substitute for the increment of G in each interval $[x_{i-1}, x_i]$ the product of the width of this interval by the derivative of G at an appropriate intermediate point t_i. Recalling that $G' = f$, we can therefore write

$$G(x_i) - G(x_{i-1}) = G'(t_i)(x_i - x_{i-1}) = f(t_i)\frac{b-a}{n},$$

where $x_{i-1} < t_i < x_i$. Thus

$$G(b) - G(a) = \frac{b-a}{n}\sum_{i=1}^{n} f(t_i).$$

Since by assumption f is integrable, passing to the limit for $n \to +\infty$ we obtain

$$G(b) - G(a) = \lim_{n\to+\infty} \frac{b-a}{n}\sum_{i=1}^{n} f(t_i) = \int_a^b f(x)\mathrm{d}x,$$

which is equation (7.7). □

Examples

4.1. We compute the area of the trapezoid determined by the function $f(x) = x^3$, $x \in [0, 1]$. An indefinite integral of $f(x) = x^3$ is $G(x) = x^4/4$ and therefore

$$\int_0^1 x^3\mathrm{d}x = \left[\frac{x^4}{4}\right]_0^1 = \frac{1}{4}.$$

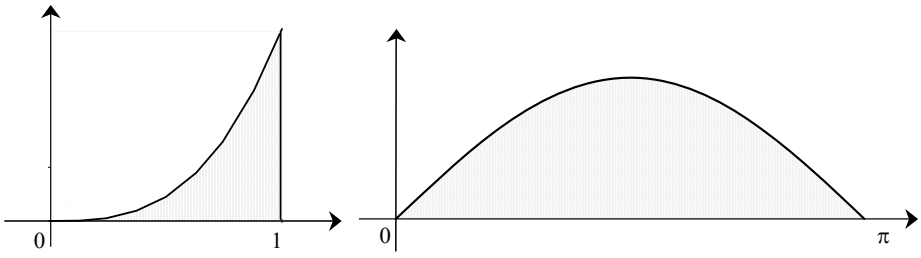


Figure 7.4. Trapezoids determined by the functions $y = x^3$ and $y = \sin x$

4.2. We compute the area of the trapezoid determined by the function $f(x) = \sin x$, $x \in [0, \pi]$. An indefinite integral of $f(x) = \sin x$ is $G(x) = -\cos x$ and therefore

$$\int_0^\pi \sin x\mathrm{d}x = [-\cos x]_0^\pi = 2.$$

4.3. With analogous calculations, we may compute

$$\int_{-1}^1 x^3\mathrm{d}x = 0, \qquad \int_0^{2\pi} \sin x\mathrm{d}x = 0.$$

The same conclusion may be obtained without calculations, by using the geometrical meaning of integral, which is the *algebraic sum of the areas of the plane regions bounded by the graph of f and the x-coordinate axis*: the areas of the plane regions above the x-coordinate axis appear in this sum with positive sign, whereas the ones below the x-coordinate axis appear with negative sign.

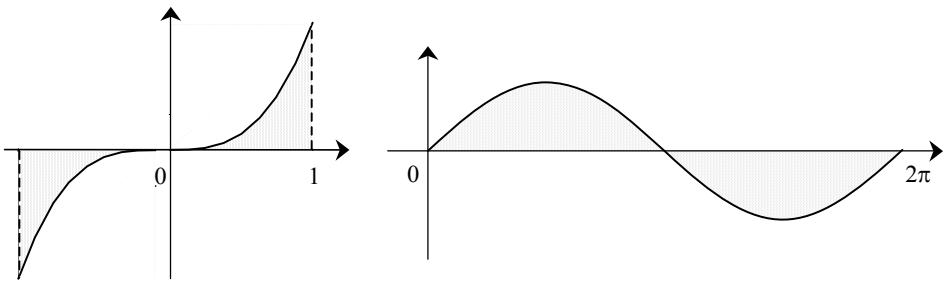


Figure 7.5. The integral as algebraic sum of areas

Formula (7.7) is important and it is useful to explain its origin and meaning. The relation between integration and derivation which appears in that formula was discovered by the mathematicians of the 17th century by looking at the relation between the position of a particle in space and its speed. Let us try to do the same... considering a travelling car.

Suppose that a car is travelling along our x-coordinate axis, and that its position at time x is described by the function $s = s(x)$. Thus, the difference $s(b) - s(a)$ shows the distance covered by the car in the time interval $[a, b]$.

Another way to describe the same movement is to consider the speed of the car. Suppose that its speed at time x is given by the function $f(x)$. Then, in an infinitesimal time interval $\mathrm{d}x$, the car will cover the distance $f(x)\mathrm{d}x$, by the physical law *space* $=$ *velocity* $\times$ *time*. We can imagine subdividing the time interval $[a, b]$ into infinitesimal intervals of length $\mathrm{d}x$ and "summing up" all the corresponding covered distances: the total covered distance will again be $s(b) - s(a)$. But the sum obtained

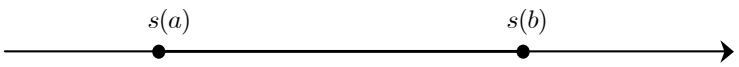


Figure 7.6.

is exactly the integral of f in $[a, b]$ and therefore we may write

$$\int_a^b f(x)\mathrm{d}x = s(b) - s(a). \tag{7.8}$$

We recall now that the derivative of s (space) w.r.t. x (time) represents the instant speed of the car:

$$s'(x) = f(x).$$

This latter formula can be read in two ways:

$$f \text{ is the derivative of } s,$$

or

$$s \text{ is an antiderivative of } f,$$

so that formula (7.8) can also be read in two ways:

from right to left: to compute the variation of s between a and b (i.e., the covered distance) we simply integrate its derivative (the speed function) in $[a, b]$;
from left to right: to compute the integral of f (the speed function) in $[a, b]$ we simply compute the variation of one of its antiderivatives (its position function) between a and b.

The arguments we illustrated in the above example are valid not only for position and speed, but for any pair of functions f, G such that $G' = f$, i.e. such that f is the derivative of G, or equivalently G is an antiderivative of f, in an interval $[a, b]$.

Let us consider some other situations, using appropriate names for the variables which appear.

• Consider a factory which produces some goods and define:
$P = P(t)$ the amount of goods produced up to the instant t;
$p = p(t)$ the amount of production in the time unit, i.e. the production speed.
Then $P'(t) = p(t)$ and the production amount during the time interval $[a, b]$ is the integral of p in $[a, b]$:

$$P(b) - P(a) = \int_a^b p(t)\mathrm{d}t.$$

• Consider again a factory which produces some goods and define:
$C = C(q)$ the total production cost of a quantity q of the considered goods;
$c = c(q)$ the marginal production cost.
Then $C'(q) = c(q)$ and the cost increment to raise production from the quantity a to the quantity b is the integral of c in $[a, b]$:

$$C(b) - C(a) = \int_a^b c(q)\mathrm{d}q.$$

Many other examples and interpretations (speed and acceleration, total and marginal utility, total and marginal profit...) are possible: we invite the reader to extend our list of examples.

7.5 The indefinite integral

The rules shown in Chapter 5 allow us to compute the derivative of nearly all elementary functions. We now want to deal with the converse problem, that is the computation of antiderivatives.

We have seen that two functions which differ by a constant have the same derivative. Is the inverse true? If we limit ourselves to functions which are defined in an interval, the answer is positive and is another consequence of Lagrange's mean value theorem.

Proposition 5.1. *Let G be differentiable in the interval (a, b), with null derivative. Then G is constant.*

Proof. We proceed as in the *monotonicity test*. Let $x_0, x \in (a, b)$, $x_0 < x$ and apply Lagrange's theorem to G in the interval $[x_0, x]$. Then there exists a point c between x_0 and x such that

$$G'(c) = \frac{G(x) - G(x_0)}{x - x_0}.$$

Since $G'(x) = 0$ for all $x \in (a, b)$, we have that $G'(c) = 0$. Therefore $G(x) - G(x_0) = 0$ and we conclude that $G(x) = G(x_0)$ for every $x \in (a, b)$, i.e., G is constant. □

As an immediate consequence, we deduce the following characterization of the family of antiderivatives of a given function.

Corollary 5.2. *Let G_1, G_2 be two antiderivatives of f in (a, b). That is,*

$$G_1'(x) = G_2'(x) \qquad \forall x \in (a, b).$$

Thus G_1 and G_2 differ by a constant, i.e., there exists a constant $c \in \mathbb{R}$ such that

$$G_1(x) = G_2(x) + c \qquad \forall x \in (a, b).$$

Proof. The function $w(x) = G_1(x) - G_2(x)$ has null derivative. Indeed:

$$w'(x) = G_1'(x) - G_2'(x) = 0$$

for all $x \in (a, b)$, therefore w is constant by theorem 5.1:

$$w(x) = G_1(x) - G_2(x) = c \text{ (constant)},$$

and $G_1(x) = G_2(x) + c$. □

We conclude that, if G is an antiderivative of f in an interval (a, b), all antiderivatives of f in the same interval are of the form $G + c$, where c is a constant.

The family of antiderivatives of f is denoted by the symbol

$$\int f(x)dx,$$

which is called the *indefinite integral* of f. Therefore, if G is an antiderivative of f in an interval (a, b), in that interval we have

$$\int f(x)\mathrm{d}x = G(x) + c, \qquad c \in \mathbb{R}.$$

Sometimes (but one should explicitly specify it), the symbol $\int f(x)\mathrm{d}x$ is also used to denote a single antiderivative.

Let us consider some examples on the computation of indefinite integrals.

Examples

5.1. Let $f(x) = e^x,\ G(x) = e^x$. Since $G'(x) = f(x)$, we have

$$\boxed{\int e^x \mathrm{d}x = e^x + c, \qquad c \in \mathbb{R}.}$$

5.2. Let $f(x) = \cos x,\ g(x) = \sin x$. Since $f'(x) = -g(x)$ and $g'(x) = f(x)$, we have

$$\boxed{\int \sin x \mathrm{d}x = -\cos x + c, \qquad \int \cos x \mathrm{d}x = \sin x + c, \qquad c \in \mathbb{R}.}$$

5.3. Let $f(x) = x^\alpha,\ \alpha \neq -1$, in $(0, +\infty)$. An antiderivative of this function is $G(x) = \dfrac{x^{\alpha+1}}{\alpha+1}$. Indeed a simple computation shows that $G'(x) = f(x)$ and thus

$$\boxed{\text{if } \alpha \neq -1, \qquad \int x^\alpha \mathrm{d}x = \frac{x^{\alpha+1}}{\alpha+1} + c, \qquad c \in \mathbb{R}.}$$

For powers defined over all of $\mathbb{R}$ (such as, for example, powers with natural exponent) the formula holds in $\mathbb{R}$.

5.4. Let $f(x) = 1/x$. An antiderivative in $(0, +\infty)$ is $G(x) = \ln x$, while in $(-\infty, 0)$ is $H(x) = \log(-x)$. We then have

$$\int \frac{1}{x}\mathrm{d}x = \begin{cases} \ln x + c_1 & \text{if } x > 0 \\ \\ \ln(-x) + c_2 & \text{if } x < 0 \end{cases} \qquad c_1, c_2 \in \mathbb{R},$$

which we sometimes summarize, without *a priori* specifying the sign of x , as

$$\int \frac{1}{x}\mathrm{d}x = \ln|x| + c,$$

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

5.5. Let $f(x) = \dfrac{1}{1+x^2}$. An antiderivative is $G(x) = \tan^{-1} x$. Therefore

$$\int \frac{1}{1+x^2}\mathrm{d}x = \tan^{-1} x + c.$$

Observing the results of examples 1-5, we may think that an antiderivative of an *elementary* function is again always an *elementary* function: the indefinite integral of an exponential is an exponential, that of the sine or cosine function is again of this type, that of the power is again a power (or, in the worst case, a logarithm), etc. Moreover, this opinion could be strengthened by the fact that *the derivative of an*

elementary function is always an elementary function (**true!**) and by the fact that, after all, taking antiderivatives is just... going back!

But this is not true: there exist important elementary functions whose antiderivatives do not admit an elementary analytical expression to represent them. Among these, for example, we have the functions:

$$f(x) = e^{-x^2}, \quad g(x) = \frac{1}{\ln x}, \quad h(x) = \frac{\sin x}{x},$$

whose antiderivatives, as we shall see, may be expressed only through a Riemann integral, which is not an elementary operation.

7.5.1 Linearity and the decomposition method

To compute indefinite integrals, we can recall the linearity property of derivatives, which implies:

$$\begin{aligned} \int [f(x) + g(x)]\,\mathrm{d}x &= \int f(x)\,\mathrm{d}x + \int g(x)\,\mathrm{d}x; \\ \int k \cdot f(x)\,\mathrm{d}x &= k \int f(x)\,\mathrm{d}x, \quad k \in \mathbb{R}. \end{aligned}$$

Examples

5.6. We compute the indefinite integral of the function

$$f(x) = 3x + 5\sqrt{x} - \frac{2}{x}, \quad x > 0.$$

We have

$$\begin{aligned} \int f(x)\,\mathrm{d}x &= 3\int x\mathrm{d}x + 5\int x^{1/2}\mathrm{d}x - 2\int \frac{1}{x}\mathrm{d}x = \\ &= 3\frac{x^2}{2} + 5\frac{2}{3}x^{3/2} - 2\ln x + c \\ &= \frac{3}{2}x^2 + \frac{10}{3}\sqrt{x^3} - 2\ln x + c. \end{aligned}$$

In the above example we used the so-called *decomposition method*. It simply consists of writing a function as a sum (more generally: a linear combination) of two or more functions, in order to compute its indefinite integral as a sum of "easier" ones. This method is typically used to compute the indefinite integral of a *rational function*, that is a quotient of polynomials.

5.7. We compute the indefinite integral of the function $f(x) = \dfrac{x^3+2}{x-1}$. We divide the numerator by the denominator[4], obtaining

$$f(x) = x^2 + x + 1 + \frac{3}{x-1}.$$

[4]Recall the identity $\dfrac{a}{b} = q + \dfrac{r}{b}$, where q and r are respectively the quotient and the remainder of the division of a by b.

We therefore have[5]

$$\int f(x)\mathrm{d}x = \frac{x^3}{3} + \frac{x^2}{2} + x + 3\ln|x-1| + c.$$

5.8. We compute

$$\int \frac{1}{x\,(M-x)}\mathrm{d}x$$

The fraction $\dfrac{1}{x\,(M-x)}$ can be obtained as the sum of two fractions of types A/x and $B/\,(M-x)$, where A, B are specific constants to be determined. Let us determine these constants A, B: we want the following equation to hold *identically*

$$\frac{A}{x} + \frac{B}{M-x} = \frac{1}{x\,(M-x)}.$$

Thus

$$A\,(M-x) + Bx = 1, \text{ that is } (-A+B)\,x + AM = 1.$$

Since this equation holds for all x, the coefficients of the terms in any degree on both sides must be equal:

$$\left\{\begin{array}{l} -A+B=0 \\ AM=1 \end{array}\right. \text{, that is } \left\{\begin{array}{l} A=1/M \\ B=1/M. \end{array}\right.$$

We thus obtain

$$\int \frac{\mathrm{d}x}{x\,(M-x)} = \int \frac{\mathrm{d}x}{Mx} + \int \frac{\mathrm{d}x}{M\,(M-x)} = \frac{1}{M}\left[\ln|x| - \ln|M-x|\right] + c$$

from which we deduce

$$\int \frac{\mathrm{d}x}{x\,(M-x)} = \frac{1}{M}\ln\left|\frac{x}{M-x}\right| + c.$$

This result is useful in many economical applications of the so-called *logistic model* (sometimes also called the *model of the S-shaped curves*).

7.5.2 Integration by parts

From the rule for the derivative of the product of two functions, we can deduce another rule for indefinite integrals, called the *rule of integration by parts*.

If f and g are differentiable functions in an interval (a, b), then

$$[fg]' = f'g + fg',$$

[5] Since we did not specify the interval in which we are searching for the indefinite integral of f, we write here the absolute value of the argument of the logarithm. The written formula is true in each of the intervals $(-\infty, 1)$ and $(1, +\infty)$.

from which we obtain

$$fg' = [fg]' - f'g.$$

By performing an indefinite integration on both sides and applying the additive property, we deduce the following formula:

$$\boxed{\int f(x)g'(x)dx = f(x)g(x) - \int f'(x)g(x)dx.} \tag{7.9}$$

The factor f appearing in the formula is called the *finite factor*, while g' is called the *differential factor*. Equation (7.9) may also be formulated in the shorter form

$$\int f\mathrm{d}g = fg - \int g\mathrm{d}f.$$

The formula of integration by parts is useful for computing indefinite integrals of functions which are products of elementary functions.

Examples

5.9. We apply formula (7.9) to compute the indefinite integral of $h(x) = xe^x$: we consider x as the finite factor and e^x as the differential factor. We obtain

$$\int xe^x\mathrm{d}x = xe^x - \int e^x\mathrm{d}x = xe^x - e^x + c.$$

What is the result if we choose e^x as the finite factor and x as the differential one? Nothing is wrong, it just means we end up with some useless calculations. Indeed we would obtain an indefinite integral which is more difficult to solve than the first one:

$$\int xe^x\mathrm{d}x = \frac{x^2}{2}e^x - \int \frac{x^2}{2}e^x\mathrm{d}x + c.$$

When choosing the finite factor and the differential factor, in case both factors have an elementary indefinite integral, it is advisable to think about which one has the simpler derivative, so that the right hand side of formula (7.9) takes the simpler form.

5.10. We compute the indefinite integral of $f(x) = \ln x$. In order to apply (7.9), we write $\ln x = 1 \cdot \ln x$ and we consider 1 as the differential factor and $\ln x$ as the finite factor (here only this choice works out). We have:

$$\int \ln x\mathrm{d}x = x\ln x - \int x \cdot \frac{1}{x}\mathrm{d}x = x\ln x - x + c.$$

7.5.3 Integration by substitution

Just as integration by parts is nothing but the rule for the derivative of a product, the method of *integration by substitution* (also called *integration by change of variable*) is precisely the rule for the derivative of a composite function.

Before showing the general formula, it is better to "warm-up" with some examples.

Examples

5.11. We compute

$$\int \cos 3x \mathrm{d}x.$$

We observe that the derivative of $\sin 3x$ is equal to $3\cos 3x$, thus we may multiply and divide by 3, obtaining

$$\int \cos 3x \mathrm{d}x = \frac{1}{3}\int 3\cos 3x \mathrm{d}x = \frac{1}{3}\sin 3x + c.$$

5.12. We compute the indefinite integral of $f(x) = \tan x$. We write $\tan x = \dfrac{\sin x}{\cos x}$ and note that the derivative of the denominator is $-\sin x$.
Now, the formula for the derivative of the logarithm of a function:

$$\frac{\mathrm{d}\ln|f(x)|}{\mathrm{d}x} = \frac{f'(x)}{f(x)}$$

(notice that the numerator is the derivative of the denominator) implies that

$$\boxed{\int \frac{f'(x)}{f(x)}\mathrm{d}x = \ln|f(x)| + c, \qquad c \in \mathbb{R}.} \tag{7.10}$$

Applying formula (7.10) in our case ($f(x) = \cos x$), we obtain

$$\int \tan x \mathrm{d}x = -\int \frac{-\sin x}{\cos x}\mathrm{d}x = -\ln|\cos x| + c, \qquad c \in \mathbb{R}$$

which is valid in each interval where $\cos x$ has a constant sign.

- *Elasticity.* The formula for the elasticity $E_f(x)$ of a function f is

$$E_f(x) = x\frac{f'(x)}{f(x)} = x\frac{\mathrm{d}\ln f(x)}{\mathrm{d}x}.$$

Calculations similar to (7.10) allow us to determine f from its elasticity. Indeed, for $x > 0$ and $f(x) > 0$, we obtain

$$\frac{\mathrm{d}\ln f(x)}{\mathrm{d}x} = \frac{E_f(x)}{x},$$

$$\ln f(x) = \int \frac{E_f(x)}{x}\mathrm{d}x$$

and finally

$$f(x) = e^{\int E_f(x)\mathrm{d}x/x}.$$

Naturally, as usual, the indefinite integral is defined up to an *additive* constant, and thus the function f is determined by its elasticity up to a *multiplicative* constant.

Let us now deal with the general method of *integration by substitution*.

The method is based on a simple remark: *If $G(x)$ is an antiderivative of $f(x)$ and $x = \phi(t)$ is a differentiable function, then the composite function $H(t) = G[\phi(t)]$ is an antiderivative of $f[\phi(t)]\phi'(t)$. That is:*

$$\boxed{\int f(x)\mathrm{d}x\Big|_{x=\phi(t)} = \int f[\phi(t)]\phi'(t)\mathrm{d}t.}$$

Indeed, since $G'(x) = f(x)$, the theorem describing the derivative of a composite function implies that

$$\frac{\mathrm{d}}{\mathrm{d}t}G[\phi(t)] = G'[\phi(t)] \cdot \phi'(t) = f[\phi(t)]\phi'(t).$$

The above boxed formula is a first substitution rule and allows us to compute indefinite integrals of all functions which can be written in the form $f[\phi(t)]\phi'(t)$, for an appropriate differentiable function ϕ: by setting $x = \phi(t)$, their indefinite integral can be obtained from the indefinite integral of $f(x)$ by substituting x with $\phi(x)$ in the result.

Suppose now that our goal is to compute an antiderivative $G(x)$ of $f(x)$ and that, while this seems a difficult task, it seems easier to compute an antiderivative $H(t)$ of $f[\phi(t)]\phi'(t)$.

If we choose the function $x = \phi(t)$ to be *invertible*, with an easily computable inverse $t = \psi(x)$, by substituting $t = \psi(x)$ in $H(t)$ we obtain $G(x)$. Indeed, since

$$\phi[\psi(x)] = x,$$

we have

$$H[\psi(x)] = G[\phi(\psi(x))] = G(x).$$

We can now obtain the following formula, which is called the formula of *integration by substitution*

$$\boxed{\int f(x)\mathrm{d}x = \int f[\phi(t)]\phi'(t)\mathrm{d}t\Big|_{t=\psi(x)}} \tag{7.11}$$

We note that, formally, we pass from one integral to the other by setting $x = \phi(t)$ and, "consequently", $\mathrm{d}x = \phi'(t)\mathrm{d}t$. This clarifies why we inserted the symbol $\mathrm{d}x$ in the symbol for the integral: when we change variable, $\mathrm{d}x$ is considered as if it really were a differential.

Once the indefinite integral

$$\int f[\phi(t)]\phi'(t)dt,$$

is computed, we obtain an expression which depends on x with the inverse substitution $t = \psi(x)$.

Usually, it is more direct to "see" the substitution $t = \psi(x)$ than the substitution $x = \phi(t)$, as the following examples show.

Examples

5.13. We compute

$$\int \frac{1}{e^x + e^{-x}} \mathrm{d}x,$$

which does not appear to be easy. Setting $t = \psi(x) = e^x$, we obtain $x = \phi(t) = \ln t$ and thus $\mathrm{d}x = \dfrac{1}{t}\mathrm{d}t$. We then have

$$\begin{aligned} \int \frac{1}{e^x + e^{-x}} \mathrm{d}x &= \int \frac{1}{t + t^{-1}} \frac{1}{t} \mathrm{d}t \Big|_{t=e^x} = \\ &= \int \frac{1}{t^2 + 1} \mathrm{d}t \Big|_{t=e^x} = \arctan e^x + c. \end{aligned}$$

5.14. We compute the indefinite integral of the function $f(x) = \cos\sqrt{x}$. Setting $t = \psi(x) = \sqrt{x}$, we obtain $x = \phi(t) = t^2,\ t \geq 0$, $\mathrm{d}x = 2t\mathrm{d}t$. We then have

$$\int \cos\sqrt{x} \mathrm{d}x = \int \cos t \cdot 2t \mathrm{d}t \Big|_{t=\sqrt{x}}.$$

Integrating by parts, we get

$$\begin{aligned} \int \cos\sqrt{x} \mathrm{d}x &= 2(t \sin t - \int \sin t \mathrm{d}t) \Big|_{t=\sqrt{x}} = 2(t \sin t + \cos t)|_{t=\sqrt{x}} + c = \\ &= 2\left(\sqrt{x} \sin\sqrt{x} + \cos\sqrt{x}\right) + c. \end{aligned}$$

Formula (7.11) can also be directly applied to the computation of Riemann integrals. If ϕ is invertible and differentiable, with a continuous derivative, we can deduce that ϕ realises a *one-to-one correspondence* between the intervals $[\alpha, \beta]$ and $[a, b]$. Denoting by ψ the inverse function of ϕ and setting $x = \phi(t)$, using formula (7.11) we obtain the following formula for Riemann integrals:

$$\int_a^b f(x)\,\mathrm{d}x = \int_{\psi(a)}^{\psi(b)} f[\phi(t)]\,\phi'(t)\mathrm{d}t. \tag{7.12}$$

The integration extrema a and b are therefore replaced by $\psi(a)$ and $\psi(b)$ respectively. If $\phi' > 0$, $\psi(a) = \alpha$ and $\psi(b) = \beta$; while if $\phi' < 0$, $\psi(a) = \beta$ and $\psi(b) = \alpha$.

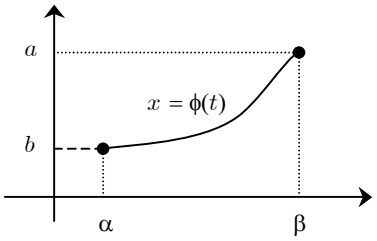


Figure 7.7.

Examples

5.15. We compute $\int_0^8 e^{\sqrt[3]{x}}\mathrm{d}x$. By setting $t = \psi(x) = \sqrt[3]{x}$ (invertible), we obtain $x = \phi(t) = t^3$ (differentiable with a continuous derivative) and $\mathrm{d}x = 3t^2\mathrm{d}t$. We thus have

$$\begin{aligned}\int_0^8 e^{\sqrt[3]{x}}\mathrm{d}x &= \int_0^2 e^t \cdot 3t^2\mathrm{d}t = \left[3t^2e^t\right]_0^2 - 6\int_0^2 te^t\mathrm{d}t = \\ &= 12e^2 - 6\left[te^t - e^t\right]_0^2 = 6e^2 - 6.\end{aligned}$$

5.16. We compute $\int_0^1 \frac{\mathrm{d}x}{e^x+1}$. We set $t = e^x$, obtaining $x = \ln t$ and $\mathrm{d}x = \frac{1}{t}\mathrm{d}t$. We thus have

$$\begin{aligned}\int_0^1 \frac{\mathrm{d}x}{e^x+1} &= \int_1^e \frac{1}{t+1}\cdot\frac{1}{t}\mathrm{d}t = \int_1^e \left(\frac{1}{t} - \frac{1}{t+1}\right)\mathrm{d}t = \\ &= \left[\ln t - \ln(t+1)\right]_1^e = 1 - \ln(e+1) + \ln 2.\end{aligned}$$

• *Change of scale.* The change of variable $x = at$ may be interpreted as a change of scale (unit of measure). For example, if x is measured in meters and we set $x = 0.01t$, then t is measured in centimeters (1 cm = 0.01 m). Setting $x = at$, if t varies between 0 and 1 then x varies between 0 and a. This fact reflects in the integral, where the symbol $\mathrm{d}x$ becomes $a\mathrm{d}t$. Thus, in the case of a linear change of scale, the conversion factor a is the same at every point of the integration interval.

If we set $x = \phi(t)$ where ϕ is no longer linear, formula (7.12) shows that $\mathrm{d}x$ becomes $\phi'(t)\mathrm{d}t$ (the best local linear approximation). This can be reinterpreted as a change of scale, where the conversion factor $\phi'(t)$ now varies from point to point in the integration interval.

7.6 Improper integrals

7.6.1 Preliminary considerations

If we want to summarize all we have said up to now about integrability conditions, we can say that a function f is certainly integrable if *its graph lies in a rectangle with*

its sides parallel to the coordinate axes, and is not too irregular therein. To say that the graph of f lies in a rectangle means that the integration interval is bounded, that is, it is an interval $[a, b]$ with $-\infty < a < b < +\infty$, and that the values of the function f on $[a, b]$ satisfy a boundedness condition $H \leq f(x) \leq K$. However, in applications we frequently meet situations in which we would like to integrate functions

— over an *unbounded* interval; for example the interval $(-\infty, b]$, which would give us an "integral" such as

$$\int_{-\infty}^{b} f(x)\mathrm{d}x.$$

— over a bounded interval, on which the function f is *unbounded*; for example, functions such as $f(x) = 1/\sqrt{x}$ on an interval $(0, b)$: note that $\lim_{x \to 0^+} f(x) = +\infty$.

The first case appears in many statistical models for the elaboration of observed data. Consider the case of an insurance company which sells a lot of insurance policies against damages. Some important results in the Calculus of Probabilities guarantee that the total amount of damage compensations that the company will have to pay, an amount which is not foreseeable with certainty, may assume different values with probabilities which are very well approximated (in the sense explained below) by the curve in Figure 8, known as a *Gaussian bell*. This curve is the graph of a Gauss function, which depends on two parameters m and σ and is given by the formula

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-m)^2/(2\sigma^2)}.$$

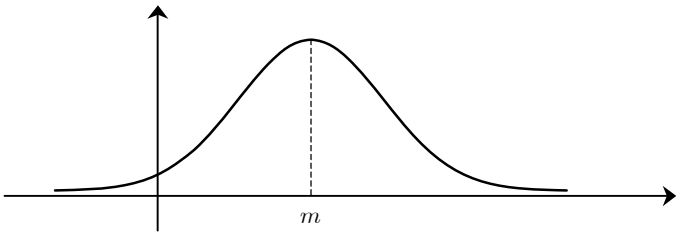


Figure 7.8. The Gauss function

The parameters m and σ, which identify the curve, are usually known to the company, therefore the curve can be considered as given. The area below the Gaussian bell is always equal to 1 and the probability that the damage compensations lie in an interval $[a, b]$ numerically corresponds to the area below the curve which is delimited by that interval.

If the insurance company has a fund a to cover damage compensations, whenever this fund is insufficient the company will enter into a condition of insolvency. Thus the fund for compensations is rather important not only for the company but also for control authorities. What is the probability that this fund is insufficient? This probability is the area under the Gaussian bell from a onwards.

We are tempted to write

$$\int_a^{+\infty} \phi(x)\mathrm{d}x, \tag{7.13}$$

but this "integral" is not covered by the theory we have presented up to now, since the integration interval $[a, +\infty)$ is not bounded.

It is also clear that normally, in order to cover the compensations, only an amount $b < a$ of the fund will be necessary, and the company will remain with the amount $a - b$. The probability that such an amount will remain to the company is the probability that the compensations will not exceed b, which is

$$\int_{-\infty}^{b} \phi(x)\mathrm{d}x, \tag{7.14}$$

and this is another "integral" which is still beyond our reach.

We have said before that the area below the graph of f is 1, corresponding to a 100% probability: this should be translated into the relation

$$\int_{-\infty}^{+\infty} \phi(x)\mathrm{d}x = 1 \tag{7.15}$$

which again involves an integration over an unbounded interval, this time unbounded both from below and from above.

Integrals such as the ones we have seen in the last three formulae are called *improper integrals* (or *generalized integrals*) and, as we shall see, it is possible to give a natural definition of them, as the limit of some Riemann integrals over bounded intervals.

7.6.2 Integrals over unbounded intervals

We start with a function f which is defined in an unbounded interval of type $[a, +\infty)$ and we suppose that this function is integrable in each interval of type $[a, k]$ $(k > a)$. To define (and also to compute)

$$I = \int_a^{+\infty} f(x)\mathrm{d}x$$

we start from the integral

$$I_k = \int_a^{k} f(x)\mathrm{d}x,$$

obtained by ending the integration interval at k (see Figure 9). The symbol I_k points out its dependency on k. We then compute

$$\lim_{k\to+\infty} I_k = \lim_{k+\infty} \int_a^{k} f(x)\mathrm{d}x. \tag{7.16}$$

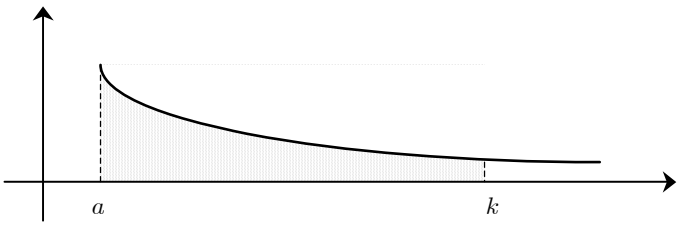


Figure 7.9.

Definition 6.1. *If the limit (7.16) exists and is finite, f is said to be* ***integrable in a generalized sense*** *in the interval $[a, +\infty)$; the value of the limit (7.16) is called the* ***improper integral*** *(or generalized integral) of f in $[a, +\infty)$. We write*

$$\int_a^{+\infty} f(x)\mathrm{d}x := \lim_{k\to+\infty} \int_a^k f(x)\mathrm{d}x.$$

We inherit the terminology we used for sequences and series, and say that the integral

$$\int_a^{+\infty} f(x)dx$$

converges if the limit (7.16) exists and is finite; it *diverges* if the limit exists but it is $+\infty$ or $-\infty$; *does not exist* if the limit does not exist.
In an analogous way, we define the improper integral in an interval of type $(-\infty, b]$ of a function f which is integrable in each interval of type $[h, b]$ with $h < b$: we start by computing

$$\int_h^b f(x)\mathrm{d}x$$

and then we set

$$\int_{-\infty}^b f(x)\mathrm{d}x := \lim_{h\to-\infty} \int_h^b f(x)\mathrm{d}x,$$

under the condition that this limit exists and is finite. We adopt the same terminology we used in the previous case.

In order to deal with integrals over the whole real axis, that is with integrals in the interval $(-\infty, +\infty)$ of a function f which is integrable in each interval of type $[h, k]$, we have to choose an arbitrary point c and then consider the two following improper integrals *separately* (this is important!)

$$\int_{-\infty}^c f(x)\mathrm{d}x = \lim_{h\to-\infty} \int_h^c f(x)\mathrm{d}x \quad \text{and} \quad \int_c^{+\infty} f(x)\mathrm{d}x = \lim_{k\to+\infty} \int_c^k f(x)\mathrm{d}x.$$

Definition 6.2. *If* ***both*** *the improper integrals considered above converge, we say that f is* ***integrable in a generalized sense*** *in $(-\infty, +\infty)$ and we set*

$$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x := \int_{-\infty}^c f(x)\mathrm{d}x + \int_c^{+\infty} f(x)\mathrm{d}x.$$

If one of the two integrals on the right hand side converges but the other diverges, or if both diverge to the same infinity, we say that the integral on the left hand side *diverges* to that infinity. In all other cases we say that the integral *does not exist*. It is easy to see that the above definition does not depend on the choice of c.

Examples

6.1. We compute $\displaystyle\int_{-\infty}^{0} e^x \mathrm{d}x$. The function $f(x) = e^x$ is continuous on the entire real axis and, therefore, it is integrable in intervals of type $[k, 0]$. We have

$$\lim_{k\to-\infty} \int_k^0 e^x \mathrm{d}x = \lim_{k\to-\infty} [e^x]_k^0 = \lim_{k\to-\infty} (1 - e^k) = 1$$

and therefore $\displaystyle\int_{-\infty}^{0} e^x \mathrm{d}x = 1$.

6.2 (*Important*). We compute, depending on the parameter $\alpha > 0$, the integral

$$\int_1^{+\infty} \frac{1}{x^\alpha} \mathrm{d}x.$$

Since the function to be integrated is continuous, it is integrable in all intervals of type $[1, k]$. If $\alpha = 1$, we have

$$\lim_{k\to+\infty} \int_1^k \frac{1}{x} \mathrm{d}x = \lim_{k\to+\infty} [\ln x]_1^k = \lim_{k\to+\infty} \ln k = +\infty,$$

and the integral diverges. If $\alpha \neq 1$, we have

$$\begin{aligned} \lim_{k\to+\infty} \int_1^k x^{-\alpha} \mathrm{d}x &= \lim_{k\to+\infty} \left[\frac{1}{1-\alpha} x^{1-\alpha} \right]_1^k = \\ &= \lim_{k\to+\infty} \frac{1}{1-\alpha} (k^{1-\alpha} - 1) = \begin{cases} +\infty & \text{if } \alpha < 1 \\ \dfrac{1}{\alpha - 1} & \text{if } \alpha > 1. \end{cases} \end{aligned}$$

Summarising:

$$\int_1^{+\infty} \frac{1}{x^\alpha} \mathrm{d}x = \begin{cases} \text{diverges to } +\infty & \text{if } \alpha \leq 1 \\ \text{converges to } \dfrac{1}{\alpha - 1} & \text{if } \alpha > 1. \end{cases} \tag{7.17}$$

6.3. We compute $\displaystyle\int_{-\infty}^{+\infty} x e^{-x^2/2} \mathrm{d}x$.

The function to be integrated is continuous on the entire real axis. By choosing $c = 0$ as the intermediate point, we have

$$\int_{-\infty}^{+\infty} x e^{-x^2/2} \mathrm{d}x = \int_{-\infty}^{0} x e^{-x^2/2} \mathrm{d}x + \int_0^{+\infty} x e^{-x^2/2} \mathrm{d}x.$$

An indefinite integral of $xe^{-x^2/2}$ is $-e^{-x^2/2}$ and therefore

$$\begin{cases} \displaystyle\int_{-\infty}^{0} xe^{-x^2/2}\mathrm{d}x = \lim_{h\to-\infty}\left[-e^{-x^2/2}\right]_h^0 = \left(-1+e^{-h^2/2}\right) = -1 \\ \displaystyle\int_{0}^{+\infty} xe^{-x^2/2}\mathrm{d}x = \lim_{k\to+\infty}\left[-e^{-x^2/2}\right]_0^k = \left(-e^{-k^2/2}+1\right) = +1. \end{cases}$$

We conclude that

$$\int_{-\infty}^{+\infty} xe^{-x^2/2}\mathrm{d}x = 0.$$

We insist in remarking that the separate study of the two improper integrals cannot be replaced with "tricks" such as

$$\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = \lim_{k\to+\infty}\int_{-k}^{+k} f(x)\,\mathrm{d}x, \tag{7.18}$$

where only *one limit* is computed[6]. Moreover, notice that the choice of linking the two extrema $-k, k$ in formula (7.18) in a special way is totally arbitrary, as well as any other choice: our next example shows indeed that the result may vary according to this choice.

6.4. We compute

$$\int_{-\infty}^{+\infty} \frac{2x}{1+x^2}\mathrm{d}x.$$

For reasons of symmetry (the function to be integrated is odd) we can deduce that

$$\int_{-k}^{+k} \frac{2x}{1+x^2}\mathrm{d}x = 0$$

for each $k > 0$. Thus

$$\lim_{k\to+\infty}\int_{-k}^{+k} \frac{2x}{1+x^2}\mathrm{d}x = 0$$

and we are tempted to write

$$\int_{-\infty}^{+\infty} \frac{2x}{1+x^2}\mathrm{d}x = 0.$$

But, as we warned above, we must examine the integrals along the two half-axis separately:

$$\int_{0}^{+\infty} \frac{2x}{1+x^2}\mathrm{d}x = \lim_{k\to+\infty}\left[\ln\left(1+x^2\right)\right]_0^k = +\infty$$

[6] Formula (7.18) actually defines another type of improper integral, called *Cauchy's principal value* and often denoted by the symbol (p.v.)$\int_{-\infty}^{+\infty} f(x)\mathrm{d}x$. It is used in important topics in probability and statistics.

and

$$\int_{-\infty}^{0} \frac{2x}{1+x^2}\mathrm{d}x = \lim_{k\to+\infty} \left[\ln\left(1+x^2\right)\right]_{-k}^{0} = -\infty.$$

Since the two improper integrals are diverging to infinity, with different signs, the required improper integral does not exist.

7.6.3 Bounded intervals and unbounded functions

Using the same method we can define the integral of an unbounded function in a bounded interval. We fix our attention on the typical case of a function which is defined in an interval (a, b), unbounded in a (right) neighbourhood of a and integrable in each interval of type $(a+\varepsilon, b)$, where $\varepsilon > 0$. To define (and compute)

$$I = \int_a^b f(x)\mathrm{d}x,$$

we start by computing the integral

$$I_\varepsilon = \int_{a+\varepsilon}^b f(x)\mathrm{d}x,$$

obtained by cutting the interval of integration at $a+\varepsilon$ (see Figure 10). The symbol I_ε underlines the dependency of the integral from ε. We then compute

$$\lim_{\varepsilon\to 0^+} I_\varepsilon = \lim_{\varepsilon\to 0^+} \int_{a+\varepsilon}^b f(x)\mathrm{d}x. \tag{7.19}$$

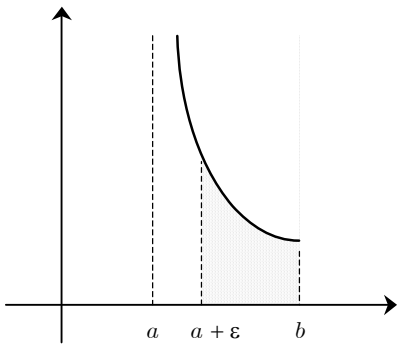


Figure 7.10.

Definition 6.3. *If the limit* (7.19) *exists and is finite,* f *is said to be* ***integrable in a generalized sense*** *in* (a, b). *The value of the limit* (7.19) *is called the* ***improper***

integral *(or generalized integral) of f in (a,b) and is denoted by*[7]

$$\int_a^b f(x)\mathrm{d}x := \lim_{\varepsilon\to 0^+}\int_{a+\varepsilon}^b f(x)\mathrm{d}x.$$

Again, if the limit (7.19) exists and is finite we say that the integral $\int_a^b f(x)\mathrm{d}x$ *converges*; if the limit is $+\infty$ or $-\infty$ we say that the integral *diverges*; if the limit does not exist then the integral *does not exist*. In a similar way, we define

$$\int_a^b f(x)\mathrm{d}x := \lim_{\varepsilon\to 0^+}\int_a^{b-\varepsilon} f(x)\mathrm{d}x$$

for functions which are defined in an interval (a,b), unbounded in a (left) neighbourhood of b and integrable in each interval $(a,b-\varepsilon)$, where $\varepsilon>0$. Finally, if f is not bounded in a neighbourhood of an internal point c (that is, a point $c\in(a,b)$), we define

$$\int_a^b f\left(x\right)\mathrm{d}x := \int_a^c f\left(x\right)\mathrm{d}x + \int_c^b f\left(x\right)\mathrm{d}x$$

under the condition that *both* integrals on the right hand side, *taken separately*, converge. If one of the two integrals on the right hand side converges and the other diverges, or if both of them diverge to the same infinity, then we say that the integral *diverges* to that infinity. In all other cases we say that the integral *does not exist*.

The previous definition extends in a natural way to the case when f is not bounded in a neighbourhood of several points of $[a,b]$.

Examples

6.5 (*Important*). Computations similar to those in example 6.2 show that

$$\int_0^1 \frac{1}{x^\alpha}\mathrm{d}x = \begin{cases} \text{diverges to } +\infty & \text{if } \alpha\geq 1 \\ \text{converges to } \dfrac{1}{1-\alpha} & \text{if } \alpha<1 \end{cases}.$$

Indeed $f(x)=\dfrac{1}{x^\alpha}$ is continuous and therefore integrable in $(\varepsilon,1)$ for every $\varepsilon>0$. In the case when $\alpha\neq 1$ we have

$$\lim_{\varepsilon\to 0^+}\int_\varepsilon^1 x^{-\alpha}\mathrm{d}x = \lim_{\varepsilon\to 0^+}\frac{1}{1-\alpha}(1-\varepsilon^{1-\alpha}) = \begin{cases} +\infty & \text{if } \alpha>1 \\ \dfrac{1}{1-\alpha} & \text{if } \alpha<1 \end{cases}, \tag{7.20}$$

while in the case $\alpha=1$ we have

$$\lim_{\varepsilon\to 0^+}\int_\varepsilon^1 \frac{1}{x}dx = \lim_{\varepsilon\to 0^+}\left[\ln x\right]_\varepsilon^1 = -\lim_{\varepsilon\to 0^+}\ln\varepsilon = +\infty.$$

[7] As we see, the symbol is the same we used for the Riemann integral. Be careful, therefore, in order to avoid confusion and errors.

In general, it is possible to prove that both integrals

$$\int_a^b \frac{\mathrm{d}x}{(x-a)^\alpha} \qquad \text{and} \qquad \int_a^b \frac{\mathrm{d}x}{(b-x)^\alpha}$$

converge if and only if $\alpha < 1$.

6.6. We compute

$$\int_{-1}^{1} \frac{x\mathrm{d}x}{\sqrt{1-x^2}}.$$

Since the function to be integrated tends to infinity for both $x \to -1$ and $x \to 1$, we have to choose a point between -1 and 1, for example $c = 0$, and separately study $\int_{-1}^{0}$ and $\int_{0}^{1}$. We have

$$\begin{aligned}\int_0^1 \frac{x\mathrm{d}x}{\sqrt{1-x^2}} &= \lim_{\delta\to 0^+} \int_0^{1-\delta} \frac{x\mathrm{d}x}{\sqrt{1-x^2}} = \lim_{\delta\to 0^+} \left[-\sqrt{1-x^2}\right]_0^{1-\delta} = \\ &= \lim_{\delta\to 0^+} \left(-\sqrt{1-(1-\delta)^2}+1\right) = 1.\end{aligned}$$

Since the function to be integrated is odd we have $\int_{-1}^{0} \frac{x\mathrm{d}x}{\sqrt{1-x^2}} = -1$ and finally

$$\int_{-1}^{1} \frac{x\mathrm{d}x}{\sqrt{1-x^2}} = 0.$$

7.6.4 Unbounded intervals and unbounded functions

When both problems appear at the same time, namely unbounded intervals and unbounded functions, we follow the previous procedures, trying to "separate" the different problematic parts. Let us illustrate this concept using an example. We want to compute

$$\int_0^{+\infty} \frac{1}{x^\alpha}\mathrm{d}x, \tag{7.21}$$

where $\alpha > 0$. The integration interval is unbounded from the right and the function is unbounded from above in a right neighbourhood of 0, that is the lower extremum of the integration interval. Thus we consider the point $c = 1$ and we study the two improper integrals

$$\int_0^1 \frac{1}{x^\alpha}\mathrm{d}x \quad \text{and} \quad \int_1^{+\infty} \frac{1}{x^\alpha}\mathrm{d}x,$$

which present only one type of "problem" each.

These integrals, computed in examples 6.2 and 6.5, converge for $\alpha < 1$ and $\alpha > 1$ respectively. This means that there is no $\alpha > 0$ for which both integrals converge. It is actually easy to see that the required integral diverges to $+\infty$ for all values of $\alpha > 0$.

We note that, if we did not take care to consider the two integrals separately, we could wrongly think that the integral exists and has value

$$\frac{1}{1-\alpha} + \frac{1}{\alpha - 1} = 0.$$

This result is clearly wrong, since the function is everywhere strictly positive and therefore the area of the underlying region must be positive.

7.6.5 Properties of improper integrals

Many properties of integrals, such as linearity and additivity with respect to the integration interval, are still valid for improper integrals. The mean value theorem remains valid for improper integrals, if the integration interval is bounded and the function is continuous at the internal points of the interval.

We write, as an example, the linearity property for integrals in $[a, +\infty)$: if f and g are integrable and α, β are real numbers, then also $\alpha f + \beta g$ is integrable and

$$\int_a^{+\infty} (\alpha f + \beta g) = \alpha \int_a^{+\infty} f + \beta \int_a^{+\infty} g.$$

This equality, if correctly read and interpreted, also holds when both integrals diverge: that is, if one of the two integrals on the right hand side diverges or both diverge to the same infinity, the integral on the left hand side also diverges.

7.7 Integrability criteria

The definition of an improper integral is quite simple, after all. It reduces the whole process to that of computing a "proper" integral and a limit. We can compute a "proper" integral according to the Fundamental Theorem of Calculus, as the variation of an antiderivative of the function to be integrated, if the task of finding an antiderivative of this function turns out to be sufficiently easy. In other cases we can try to overcome this difficulty with the methods of numerical calculus (and the help of a computer). Even if they do not allow us to calculate the exact value of the integral, they give us at least a good approximation, which is usually more than enough for practical applications. Of course, before using a computer, we have to make sure that the integral converges. To this purpose, in this section we give two criteria of integrability for positive functions and a more general criterion for functions with no assumptions about the sign, fixing our attention on integrals over an unbounded interval $[a, +\infty)$: similar considerations hold in the other cases.

Proposition 7.1 (Comparison criterion). *Let f and g be two functions defined over $[a, +\infty)$, integrable in each interval $[a, k]$ with $k > a$, and satisfying the condition:*

$$\text{for every } x \in [a, +\infty) \qquad 0 \le f(x) \le g(x).$$

Then: if g is integrable in $[a, +\infty)$, f is also integrable; if f is not integrable in $[a, +\infty)$, g is also not integrable.

Proof. The monotonicity property of integrals implies that

$$0 \leq \int_a^k f \leq \int_a^k g.$$

Passing to the limit for $k \to +\infty$, the comparison theorem for limits implies the thesis. □

Example 7.1 (*Important*). Recall the Gauss curve: let us convince ourselves that the Gauss function is really integrable on the whole real axis, so that we can give a meaning to the integrals (7.13), (7.14), (7.15) which we wrote at the beginning of section 6. We consider the case $m = 0$ and $\sigma = 1$, that is the function[8]

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

We observe that the constant $1/\sqrt{2\pi}$ has no influence on the convergence of the integral. Moreover ϕ is an *even* function, continuous on all the real axis, so that it is sufficient to verify the existence of the integral

$$\int_a^{+\infty} e^{-x^2/2} \mathrm{d}x$$

where $a > 0$. We consider the two functions $f_1(x) = x$ and $f_2(x) = x^2/2$. For $x > 2$ the second function is bigger than the first one:

$$\frac{x^2}{2} > x.$$

This implies that

$$e^{-x^2/2} < e^{-x} \quad \text{for } x > 2.$$

The function $F_1(x) = e^{-x}$ is integrable in a generalized sense on each interval $(a, +\infty)$, since

$$\int_a^{+\infty} e^{-x} \mathrm{d}x = \left[-e^{-x}\right]_a^{+\infty} = e^{-a} < +\infty,$$

and therefore the same holds for $F_2(x) = e^{-x^2/2}$.

The actual calculation of the given integral requires a deeper knowledge than that which can be acquired in a general course of Mathematics. However we have already mentioned, more or less implicitly, that the area under the Gauss bell is equal to 1, that is:

$$\boxed{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-x^2/2} \mathrm{d}x = 1,}$$

from which we can deduce $\displaystyle\int_{-\infty}^{+\infty} e^{-x^2/2} \mathrm{d}x = \sqrt{2\pi}$.

[8] Without loss of generality, we can always reduce our problem to this case by the change of variable $t = (x - m)/\sigma$. This case, often used in statistics, is called the *standard Gauss function.*

As a consequence of the comparison criterion, another very useful criterion follows.

Corollary 7.2 (Asymptotic comparison criterion). *Let f and g be two positive functions defined in $[a, +\infty)$, integrable in each interval $[a, k]$ with $k > a$, and let*

$$f(x) \sim g(x) \qquad \text{as } x \to +\infty;$$

then f is integrable in $[a, +\infty)$ if and only if g is.

Proof. Since

$$\lim_{x \to +\infty} \frac{f(x)}{g(x)} = 1,$$

it follows that for $x > c$ (c large enough)

$$\frac{1}{2} < \frac{f(x)}{g(x)} < 2,$$

which gives

$$\frac{1}{2} g(x) < f(x) < 2g(x).$$

By the homogeneity property, the functions $\frac{1}{2}g$ and $2g$ are integrable if and only if g is. The thesis then follows from the comparison criterion. □

The additive property with respect to the integration interval and the homogeneity property allow us to reduce the study of an integral for which the function to be integrated has a constant sign, at least for values of the variable which are sufficiently large, to the study of an integral of a function which is always positive. For example, if f is negative for all values x which are sufficiently large ($f(x) < 0$ for $x > c$, for some $c \in \mathbb{R}$), we can write

$$\int_a^{+\infty} f(x)\,\mathrm{d}x = \int_a^c f(x)\mathrm{d}x + \int_c^{+\infty} f(x)\mathrm{d}x$$

and then we can limit ourselves to studying the integrability of $-f$ in $[c, +\infty)$.

If the function to be integrated has a sign which is not constant for all sufficiently large values, we may use the following (very weak) criterion.

Proposition 7.3 (Absolute convergence criterion). *Let $f : [a, +\infty) \to \mathbb{R}$ be an integrable function in each interval $[a, k]$ with $k > a$ and suppose that $|f|$ is integrable in $[a, +\infty)$. Then f is also integrable in $[a, +\infty)$ and*

$$\left| \int_a^{+\infty} f(x)\,\mathrm{d}x \right| \leq \int_a^{+\infty} |f(x)|\,\mathrm{d}x$$

Note the similarity between the three criterions illustrated here and the three criterions for series with the same name.

Examples

7.2. We check that $f(x) = \ln x/x^2$ is integrable in $[1, +\infty)$. Since f is continuous in $(0, +\infty)$, it is integrable in each interval $[1, k]$ with $k > 1$. For $x \geq 1$ we have, for example[9], that

$$\ln x < \sqrt{x},$$

and thus

$$\frac{\ln x}{x^2} < \frac{\sqrt{x}}{x^2} = \frac{1}{x^{3/2}}.$$

Since the function $g(x) = x^{-3/2}$ is integrable in $[1, +\infty)$ (see example 6.2), f is also integrable by the comparison criterion.

7.3. $f(x) = \dfrac{x + \ln x}{x^2 + x + e^{-x}}$ is not integrable in $[1, +\infty)$. Indeed

$$\text{for } x \to +\infty \qquad f(x) \sim \frac{1}{x}$$

and $g(x) = 1/x$ is not integrable in a neighbourhood of $+\infty$ (see example 6.2).

7.4. $f(x) = \sin x/x^3$ is integrable in $[1, +\infty)$. We observe that f is continuous and integrable in every interval $[1, k]$ with $k > 1$, but its sign is not constant for all values x which are larger than a given value. Let us try with $|f(x)| = |\sin x| / x^3$.

By the comparison criterion, $|f|$ is integrable; indeed

$$\frac{|\sin x|}{x^3} \leq \frac{1}{x^3}$$

and is therefore a minorant of an integrable function (see example 6.2). By the absolute convergence criterion, we deduce that f is integrable.

7.8 Series and integrals

We have already noted a similarity between improper integrals in intervals of type $[a, +\infty)$ and numerical series. This similarity is not fortuitous. Indeed, given a regular series $\sum_{n=0}^{+\infty} a_n$ and considering the function $f : [0, +\infty) \to \mathbb{R}$ defined by

$$f(x) = \ a_n, \quad n \leq x < n + 1,$$

we can easily check that

$$\int_0^{+\infty} f(x)\mathrm{d}x = \sum_{n=0}^{+\infty} a_n.$$

That is, every regular series can be regarded as the improper integral of a suitable function. From this remark we immediately deduce that, under appropriate conditions on the functions to be integrated, integrals and series control each other. Precisely, we have the following

[9] $\ln x/\sqrt{x} \to 0$ if $x \to +\infty$.

Proposition 8.1 (Comparison criterion between series and integrals). *Let $f : [0, +\infty) \to \mathbb{R}$ be a positive and decreasing function, integrable in every interval $[0, k]$. If we define $a_n = f(n)$ for all n, we have*

$$\sum_{n=1}^{+\infty} a_n \leq \int_0^{+\infty} f(x)dx \leq \sum_{n=0}^{+\infty} a_n.$$

In particular,

$$\sum_{n=0}^{+\infty} a_n < \infty \qquad \text{if and only if} \qquad \int_0^{+\infty} f(x)\mathrm{d}x < \infty.$$

Proof. On every interval $[0, k]$, where k is a positive integer, Figure 11 shows that

$$\sum_{i=1}^{k} a_i \leq \int_0^{k} f(x)\mathrm{d}x \leq \sum_{i=0}^{k-1} a_i.$$

By the comparison theorem for limits of sequences, passing to the limit for $k \to +\infty$ we deduce the thesis. □

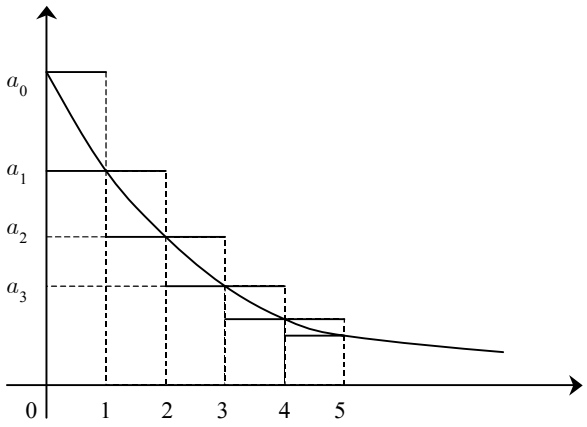


Figure 7.11. Comparison between series and integrals

Evidently, the above criterion can be used for studying the behaviour of a numerical series, when the calculation of the corresponding integral is easier. A typical example is given by the generalized harmonic series: recalling example 4.1 of Chapter 6 and example 6.2 of this chapter, we deduce immediately that

$$\boxed{\sum_{n=1}^{+\infty} \frac{1}{n^\alpha} \text{ converges if and only if } \alpha > 1.}$$

Example 8.1. We study

$$\sum_{n=2}^{+\infty} \frac{1}{n \ln n}.$$

Setting $f(x) = 1/(x \ln x)$, f is a positive function, decreasing and continuous in $[2, +\infty)$, thus integrable in every interval $[2, k]$. The hypotheses of theorem 8.1 are then satisfied. We compute

$$\lim_{k \to +\infty} \int_2^k \frac{\mathrm{d}x}{x \ \ln x} = \lim_{k \to +\infty} [\ln \ln x]_2^{+\infty} = \lim_{k \to +\infty} (\ln \ln k - \ln \ln 2) = +\infty.$$

Therefore f is not integrable and the given series diverges.

7.9 Integral functions

We have obtained a formula to compute Riemann integrals when the function to be integrated has a known antiderivative. We have also explicitly computed antiderivatives for some functions. On the other hand, we have also remarked that the functions for which we can write antiderivatives using elementary functions are rather few.

We recall that there are functions of remarkable importance (not only in a theoretical sense), such as, for example, $f(x) = e^{-x^2}$, which do not admit an elementary antiderivative. But do such functions have an antiderivative at all? Or, more generally, are there functions which have no antiderivative?

A partial, but very significative, answer is the following:

Every continuous function in $[a, b]$ has an antiderivative.

We can prove such a result by investigating further the link between the concepts of integral and derivative. We use once again the connection between position and speed, illustrated in the introductory section of this Chapter. Consider formula (7.8) and keep the point a fixed, looking for the space covered up to the instant x. The formula which describes the space covered is obtained by substituting x for b and changing the symbol used for the integration variable, in order to avoid confusion:

$$\int_a^x f(t)\mathrm{d}t = s(x) - s(a). \tag{7.22}$$

If s is differentiable, we know that $s'(x) = f(x)$ and therefore, differentiating both sides of (7.22), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}x} \int_a^x f(t)\mathrm{d}t = f(x). \tag{7.23}$$

Formula (7.23) tells us that, in some sense, derivative and integral are, as operations, inverses of each other. But we have to remark that formula (7.23) was obtained *by using the information that s is differentiable and that $s'(x) = f(x)$*; indeed, this information allowed us to differentiate both sides of (7.22), which have to be equal.

Thus, we leave the previous example and we perform some explicit calculations in order to see if (7.23) holds in general. We consider the function $f(t) = 2t$ and its integral between 1 and x. We have

$$\int_1^x 2t\mathrm{d}t = x^2 - 1,$$

implying

$$\frac{\mathrm{d}}{\mathrm{d}x}\int_a^x f(t)\mathrm{d}t = \frac{\mathrm{d}}{\mathrm{d}x}(x^2 - 1) = 2x = f(x).$$

We deduce that, at least in this case, formula (7.23) is true. Let us try with a function having a jump discontinuity. For example

$$f(t) = \begin{cases} -1 & \text{if } -1 \leqslant t < 0 \\ 1 & \text{if } 0 \leqslant t \leqslant 1 \end{cases}.$$

We find that (the reader can check the result using the geometrical meaning of the integral and Figure 12):

$$\int_{-1}^x f(t)\mathrm{d}t = |x| - 1.$$

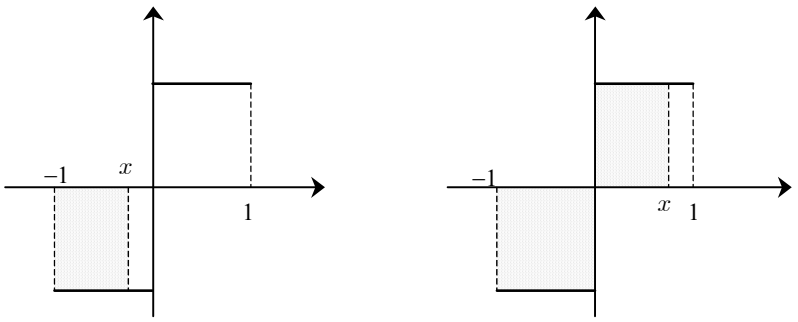


Figure 7.12.

The function $F(x) = |x| - 1$ presents a corner at the origin where, therefore, *it has no derivative!* We have discovered that a jump discontinuity for the function to be integrated caused a corner for the integral. Thus (7.23) *does not hold* for all integrable functions.

Anyway, the second example shows us the way forward: let us consider *only* continuous functions. In this case, we may still hope that (7.23) holds.

We can now formalize the preceding remarks. We consider a function f which is integrable in an interval $[a, b]$ and we consider two points c, $x \in [a, b]$. The function f is also integrable in the interval with extrema c and x ($[c, x]$ or $[x, c]$, depending on whether $c \leq x$ or $x \leq c$). Let us fix c and allow x to vary. We can associate with

each x the value of the integral[10] between c and x

$$x \longmapsto \int_c^x f(t)\mathrm{d}t;$$

this way we construct a function F_c called an *integral function*. Geometrically, if f is positive $F_c(x)$ represents the area (with sign!) hatched in Figure 13.

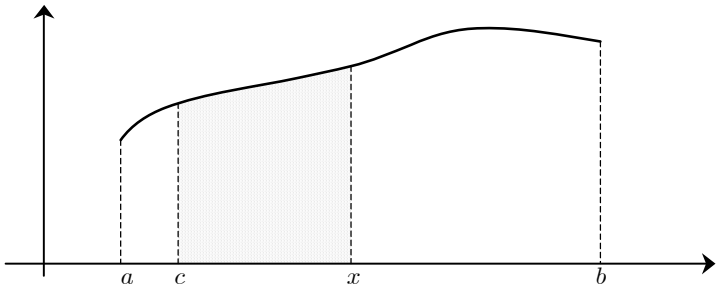


Figure 7.13. The integral function F_c is the area of the hatched region

By the way, we observe that, choosing a different point c, the integral functions differ by a constant: indeed, by the additivity property with respect to the integration interval, we have

$$F_a(x) = \int_a^x f(t)\mathrm{d}t = \int_a^c f(t)\mathrm{d}t + \int_c^x f(t)\mathrm{d}t = \int_a^c f(t)\mathrm{d}t \; + F_c(x)\,.$$

As we see, the difference between $F_a(x)$ and $F_c(x)$ is the number $\int_a^c f(t)\mathrm{d}t$, which does not depend on x.

The integral function is well defined even if the integration interval is unbounded, in the case that f is integrable in a generalized sense. Specially important is the integral function associated with the *Gauss function*, defined by choosing $c = -\infty$:

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2}\mathrm{d}t,$$

which frequently appears in statistics textbooks.

The following result, known as the *Second Fundamental Theorem of Calculus*, shows that our initial hope was well placed, when f is continuous.

Theorem 9.1. *If $f : [a, b] \to \mathbb{R}$ is continuous, then the integral function*[11]

$$F(x) = \int_a^x f(t)\,\mathrm{d}t, \qquad x \in [a, b]$$

[10] We called t the integration variable (which is a dummy variable, as we know), in order to avoid confusion with the true variable x, which appears as the upper extreme of the integral.

[11] Obviously, the thesis holds as well for every other integral function F_c.

has derivative in (a, b) *and*

$$F'(x) = f(x). \tag{7.24}$$

In other words, for a continuous function in an interval $[a, b]$ any of its integral functions is an antiderivative.

Proof. We compute the increment of F from x to $x + h$

$$F(x+h) - F(x) = \int_a^{x+h} f(t)\,\mathrm{d}t - \int_a^x f(t)\,\mathrm{d}t =$$

(by the additive property with respect to the integration integral)

$$= \int_x^{x+h} f(t)\,\mathrm{d}t.$$

We now use the mean value theorem 3.1: we obtain

$$\frac{1}{h}\int_x^{x+h} f(t)\,\mathrm{d}t = f(c),$$

for some appropriate point c lying between x and $x + h$. Therefore:

$$\frac{F(x+h) - F(x)}{h} = f(c).$$

If we pass to the limit for $h \to 0$, we have that $c \to x$ and, by the continuity of f, $f(c) \to f(x)$. The limit on the left hand side exactly defines $F'(x)$ and thus $F'(x) = f(x)$. □

The second fundamental theorem of calculus shows *how to construct an antiderivative* of a continuous function f. For example, as we have already noted, the Gauss function

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

has no elementary antiderivative. The theorem shows that, in any case,

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\mathrm{d}t$$

is an antiderivative of ϕ, which means exactly

$$\Phi'(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

7.10 Exercises

7.1. Compute

$$\int \frac{x+2}{2x-3}\mathrm{d}x, \quad \int \frac{x+1}{x^2+1}\mathrm{d}x, \quad \int \frac{x^2\mathrm{d}x}{x^2-1}, \quad \int \frac{\mathrm{d}x}{(2x+1)^2}.$$

7.2. The net investment $I(t)$ is defined as the rate of variation of the process of constructing a principal $K(t)$ over time t, i.e., when we have a continuous process over time, $I(t) = K'(t)$. If the rate of variation $I(t)$ is continuous over time, determine the function $K(t)$, knowing that the net investment is $I(t) = 40\sqrt[5]{t^3}$ and the principal at time $t = 0$ is worth 75.

7.3. The marginal production cost of some given goods is

$$c_m(q) = 25 + 30q + 9q^2,$$

while the fixed cost $c_f = c(0)$ is 55. Determine the total cost $c(q)$.

7.4. Using formula (7.9), compute an antiderivative of $f(x) = \tan^{-1} x$.

7.5. Check the following formulae, using the rule for the derivative of a composite function.

$$\int f'(x)e^{f(x)}\mathrm{d}x = e^{f(x)} + k, \qquad k \in \mathbb{R},$$

$$\int f'(x)\sin f(x)\mathrm{d}x = -\cos f(x) + k, \qquad k \in \mathbb{R},$$

$$\int f'(x)\cos f(x)\mathrm{d}x = \sin f(x) + k, \qquad k \in \mathbb{R},$$

$$\int f'(x)\,[f(x)]^\alpha\,\mathrm{d}x = \frac{[f(x)]^{\alpha+1}}{\alpha+1} + k, \qquad k \in \mathbb{R},\ \alpha \neq -1.$$

7.6. Compute

$$\int x\sqrt{x^2+1}\mathrm{d}x, \quad \int \frac{\mathrm{d}x}{x\ln x}, \quad \int \sin x\cos x\mathrm{d}x, \quad \int \frac{e^x\mathrm{d}x}{(e^x+3)^3}.$$

7.7. Write the antiderivative of the function

$$f(x) = \frac{1}{e^x+1}$$

whose graph passes through the point $(0, 0)$.

7.8. Write the antiderivative of the function

$$f(x) = \frac{\ln(x+2)}{x^2}$$

which is infinitesimal for $x \to +\infty$.

7.9. Using two different integration methods, write the indefinite integral of

$$f(x) = \frac{x}{\sqrt{x+1}}.$$

7.10. Compute

$$\int_0^1 (x+3)e^{-x}\mathrm{d}x.$$

7.11. Compute the area of the plane region bounded by two parabola arcs having equations $y = ax^2$ and $y = a\sqrt{x}$ respectively. What is the value of a if the area is exactly 12?

7.12. Compute the mean value of the function $f(x) = \cos x$ in $[-\pi/2, \pi/2]$.

7.13. *Income tax. Pareto's Model.* Suppose that the "income curve", in the relevant segment $[a, b]$, has equation

$$y = n(x) = \frac{A}{x^\alpha},$$

(with A a positive constant and $\alpha > 1$), where y is the number of individuals whose income is larger than x. The number of individuals whose income lies between x_1 and x_2 is given by

$$-\int_{x_1}^{x_2} n'(x)\mathrm{d}x = \frac{A}{x_1^\alpha} - \frac{A}{x_2^\alpha}.$$

If the average tax rate for an income x is given by $\gamma(x) = a + b\sqrt{x} + cx$, compute the income tax obtained from all individuals whose income lies between x_1 and x_2, given by

$$-\int_{x_1}^{x_2} xn'(x)\gamma(x)\,\mathrm{d}x.$$

7.14. Compute the following integrals, by using the suggested change of variable:

$$\int_1^e \frac{\ln x}{x(\ln x + 1)}\mathrm{d}x, \qquad t = \ln x.$$

$$\int_0^{\pi/2} \frac{\cos x}{\sqrt{\sin x + 2}}\mathrm{d}x, \qquad t = \sin x.$$

7.15. *Consumer surplus and producer surplus.* Let $p = f_\mathrm{s}(q)$ be a supply function, representing the prices of different quantities of some supplied goods, and let $p = f_\mathrm{d}(q)$ be a demand function[12], representing the prices which the consumers are willing to pay for different quantities of the same goods. If (q_0, p_0) is the equilibrium price in the market, the total revenue of the producers who would offer the goods at a lower price is given by

$$q_0 p_0 - \int_0^{q_0} f_\mathrm{s}(q)\mathrm{d}q$$

[12] More precisely, they are the inverse functions of the usual supply and demand functions.

and is called the producer surplus. The total benefit of the consumers who are willing to pay a price greater than p_0 is given by

$$\int_0^{q_0} f_{\mathrm{d}}(q)\mathrm{d}q - q_0 p_0$$

and is called the consumer surplus.

(a) Give a geometrical interpretation of the producer and consumer surpluses.

(b) Let $p_{\mathrm{d}} = 25 - q^2$ and $p_{\mathrm{s}} = 2q + 1$, be respectively the demand and the supply functions of some given goods. Under the hypothesis of perfect competition, determine the consumer and producer surpluses.

7.16. A factory producing some consumer goods has a profit which varies during time according to the law $\pi(t) = (10 + 2t)\, e^{-t/10}$. Supposing that the instant interest rate is $\delta = 10\%$, compute the discounted value $A\,(T)$ of all profits realized after time T:

$$A\,(T) = \int_T^{+\infty} \pi\,(t)\, e^{-\delta t}\mathrm{d}t.$$

7.17. Suppose that, in a society made up of N working individuals, the minimum income is R_0 and the number $F\,(R)$ of individuals whose income is less than R obeys Pareto's law

$$F\,(R) = N\left[1 - \left(\frac{R_0}{R}\right)^{3/2}\right]$$

(a) Show that the integral

$$\int_{R_0}^{+\infty} F'\,(R)\,\mathrm{d}R$$

converges and compute its value. Interpret this result.

(b) Compute the sum of all individual incomes which are greater than R_1. Deduce the mean value $\overline{S}\,(R_1)$ of all individual incomes which are greater than R_1 and the mean value $\overline{R}$ of the individual incomes of the whole population.

7.18. Without computing the integral, determine all possible points of local maximum and local minimum and all possible inflection points for the function

$$F\,(x) = \int_0^x \left(t^3 - t^2\right) e^t \mathrm{d}t.$$

7.19. *Find the error.* The integral of $f(x) = 1/x$ between -1 and $+1$ is 0. Indeed an antiderivative of f is $\ln|x|$ and we have

$$\int_{-1}^{1} \frac{1}{x}\mathrm{d}x = \left[\ln|x|\right]_{-1}^{1} = 0.$$

7.20. Determine whether the following integrals converge and, if they do converge, calculate them:

(a) $\int_{-\infty}^{+\infty} \dfrac{1}{x^2+1}\mathrm{d}x$, ($b$) $\int_0^2 \dfrac{\mathrm{d}x}{\sqrt{2-x}}$,

(c) $\int_2^{+\infty} \dfrac{\ln x}{x}\mathrm{d}x$, ($d$) $\int_0^{+\infty} x^2 e^{-x}\mathrm{d}x$.

7.21. Determine whether the following integrals converge:

$$(a)\ \int_1^{+\infty} \frac{\sqrt{x}+3}{x+2\ln x}\mathrm{d}x, \qquad (b)\ \int_0^{+\infty} \cos x \cdot e^{-x}\mathrm{d}x.$$

7.22. *Find the error.* The derivative of the function

$$F(x) = \int_0^{\sqrt{x}} e^{-t^2} dt$$

is $F'(x) = e^{-x}$.

7.23. Choose the correct formula, specifying if any extra hypotheses on f are necessary.

$$\int_a^b f = f'(b) - f'(a), \qquad \int_a^b f' = f(b) + f(a), \qquad \int_a^b f' = f(b) - f(a).$$

7.24. Determine the behaviour of the series

$$\sum_{n=2}^{+\infty} \frac{1}{n(\ln n)^2}.$$

8

Vectors and Matrices

Many mathematical models and applications can be expressed in terms of *vectors* and *matrices*[1]. Roughly speaking, these objects are numerical tables which can be "combined" through rules which can be very conveniently defined, resembling the operations on numbers. With these operations (an *algebra*), vectors and matrices become a very natural tool for describing various economic and social issues.

The chapter is organised as follows.

• *Vectors* are introduced as ordered n-tuples of real numbers, and the operations of sum between vectors, product of a vector by a scalar and inner product of two vectors are defined.

• The concepts of *linear dependence* and *independence*, and that of *subspace spanned* by a set of vectors, are featured.

• *Matrices* are introduced as matchings of vectors, and the operations of sum of matrices, product of a matrix by a scalar and product of two matrices are defined.

• Finally, the definitions of *determinant* (of a square matrix) and *rank* (of any matrix) are covered.

8.1 Vectors in $\mathbb{R}^n$

In social sciences, it is common to deal with objects which cannot be described by means of *a single* number but require an ordered list of them:

[1] *Matlab*, one of the most widely used softwares in recent applications of financial mathematics, was created as a tool for performing calculations on the objects which we shall introduce in this chapter.

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}. \tag{8.1}$$

Example 1.1. Suppose that a firm's warehouse contains goods of (up to) n different kinds. It is natural to describe the amount of the warehouse stock by means of a list like (8.1), where the numbers x_i are the physical amounts (in kgs, number of items, ...) of the goods of each type, numbered by $i = 1, 2, \ldots, n$. If, for instance, the warehouse contains four types of goods, then the list

$$\begin{bmatrix} 150 \\ 30 \\ 200 \\ 0 \end{bmatrix}$$

means that there are 150 units of the first kind, 30 of the second, 200 of the third and none of the fourth.

Example 1.2. A financial operation involves the cash flows $x_1, x_2, \ldots, x_n$ at the maturities $1, 2, \ldots, n$ (where outlays bear a negative sign, and incomes a positive one). We can then describe it by means of a list such as (8.1). If, for instance, buying a zero-coupon bond today at the price 725 entitles us to collect the amount 1000 in two years, we consider the annual maturities and describe the operation as follows:

$$\begin{bmatrix} -725 \\ 0 \\ 1000 \end{bmatrix}.$$

We shall call a list like (8.1) a *column vector*. The numbers within such a list are called the *components* of the vector. We will adopt the convention of indicating the components with the same letter used to denote the vector: thus, the components of $\mathbf{x}$ will be written as $x_1, x_2, \ldots, x_n$.

Various reasons (e.g. saving space) might induce us to use *row vectors*, where the components are aligned horizontally:

$$\begin{bmatrix} x_1 & x_2 & \ldots & x_n \end{bmatrix} \tag{8.2}$$

Definition 1.1. *A* ***vector***[2] *is an ordered n-tuple of real numbers .*

We remark that (8.1) and (8.2) contain essentially the same information. Therefore, we shall call any object which is built with n ordered real numbers a *vector*, regardless of the way they are written. We shall denote the object (8.1) by $\mathbf{x}$, reserving the notation $\mathbf{x}^{\mathrm{T}}$ (*transpose*[3] vector $\mathbf{x}$) for (8.2). The notation (8.2) can be replaced by the notation $(x_1, x_2, ..., x_n)$.

[2] The term vector has indeed a more general meaning. Ordered n-tuples of real numbers can be considered as particular vectors when one defines upon them the operations we are going to consider in the following pages.

[3] The transposition operation will be formally defined in Section 7.

The set of all vectors with n real components is denoted by $\mathbb{R}^n$.

Real numbers can be thought of as single component vectors. Vectors in examples 1 and 2 are, respectively, elements of $\mathbb{R}^4$ and $\mathbb{R}^3$.

Vectors with 1, 2 or 3 components can be geometrically represented, respectively, on a line ("axis"), in a plane, and in the three-dimensional Cartesian space. This representation can be performed, depending on individual choices, with *dots* or with *arrows*[4].

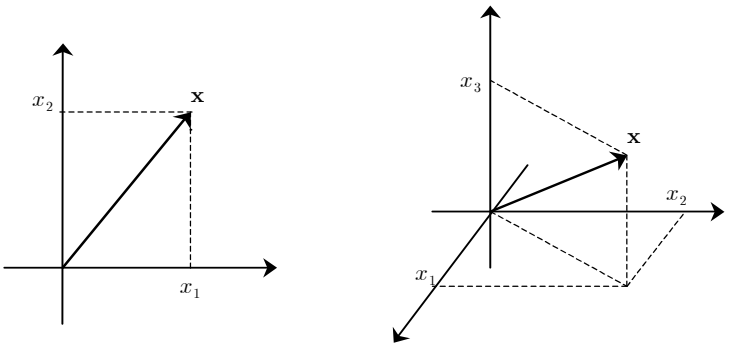


Figure 8.1. Vectors in $\mathbb{R}^2$ and in $\mathbb{R}^3$

Among the infinitely many vectors of $\mathbb{R}^n$ there are n vectors of particular importance, called the *fundamental vectors* of $\mathbb{R}^n$ and denoted by the symbols $\mathbf{e}^1, \dots \mathbf{e}^n$. Every vector $\mathbf{e}^i$, with $i = 1, \dots, n$, has all of its components equal to zero with the exception of the i-th one, which has unit value. The fundamental vectors of $\mathbb{R}^3$, for instance, are

$$\mathbf{e}^1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \qquad \mathbf{e}^2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \qquad \mathbf{e}^3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

A vector is, then, a list of numbers whose meaning is related to the order in which those numbers are listed. On vectors, a set of "calculation rules" can be assigned, which in part work like the analogous calculation rules on numbers. This arithmetic is frequently used in economic theory and constitutes the basic language for many useful calculation techniques. The most convenient feature of such a language is its "typographic" aspect: any number (possibly thousands) of operations or relations can be written as *a single operation* or *a single relation*. Since many mathematical softwares (*Matlab, Maple, Mathematica, Mathcad...*) can understand this typography, vectors and matrices are also a powerful device for translating the calculations we need into computational implementations.

[4]Representing vectors by means of arrows is surely familiar to the reader with some knowledge of Physics.

8.2 Operations with vectors

The most simple issues in the arithmetic of numbers are equalities and inequalities. With vectors, they are defined as follows.

• *Equality.* Two vectors are *equal* when they feature the same number of components and the components in the same places coincide. If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the equality $\mathbf{x} = \mathbf{y}$ is a shorthand notation for the n equalities

$$x_1 = y_1, \quad x_2 = y_2, \ldots, \quad x_n = y_n.$$

The relation of equality between vectors satisfies the same three properties as the equality between numbers:

$\mathbf{x} = \mathbf{x}$	reflexivity
if $\mathbf{x} = \mathbf{y}$, then $\mathbf{y} = \mathbf{x}$	symmetry
if $\mathbf{x} = \mathbf{y}$ and $\mathbf{y} = \mathbf{z}$, then $\mathbf{x} = \mathbf{z}$	transitivity.

We shall write: $\mathbf{x} \neq \mathbf{y}$ when the two vectors $\mathbf{x}, \mathbf{y}$ are not equal.

• *Order.* Unlike the generalization of the concept of equality, more caution is needed when defining inequalities between vectors. Indeed, given any two real numbers, either one is greater than the other or they are equal, but ordering vectors is not as simple as for numbers. A good idea might be to consider a vector $\mathbf{x}$ *greater* than the vector $\mathbf{y}$ if *all of* the components of $\mathbf{x}$ are greater than the corresponding components of $\mathbf{y}$:

$$\mathbf{x} > \mathbf{y} \quad \textit{if and only if } x_s > y_s \quad \textit{for every } s = 1, \ldots, n. \tag{8.3}$$

It is easy to realise that one can encounter situations when none of the relations $\mathbf{x} > \mathbf{y}$ or $\mathbf{y} > \mathbf{x}$ or $\mathbf{x} = \mathbf{y}$ holds, as in the case of the vectors

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Because of this, the order between vectors defined by (8.3) is called a *partial order*. This means that, given two vectors $\mathbf{x}, \mathbf{y}$, if we are lucky we can tell that $\mathbf{x} > \mathbf{y}$ or that $\mathbf{y} > \mathbf{x}$ or that $\mathbf{y} = \mathbf{x}$, but, in many cases, none of these holds. The order defined by (8.3) satisfies the so-called transitive property.

Besides the strict ordering just defined, another and less restrictive one can be analogously defined.

$$\mathbf{x} \geqq \mathbf{y} \textit{ if and only if } x_s \geq y_s \textit{ for every } s = 1, \ldots, n.$$

Of course, this is a partial order as well. By comparing the vectors in $\mathbb{R}^n$ with the zero vector $\mathbf{0}$, which has every component equal to 0, we can introduce the concepts of *positive* vector: $\mathbf{x} > \mathbf{0}$, i.e. $x_s > 0$ for every $s = 1, \ldots, n$ and *non-negative* vector: $\mathbf{x} \geqq \mathbf{0}$, i.e. $x_s \geq 0$ for every $s = 1, \ldots, n$[5].

[5] In economic theory another inequality between vectors, intermediate between the two discussed here, is sometimes needed. One says that $\mathbf{x}$ is *weakly greater* than $\mathbf{y}$ ($\mathbf{x} \geq \mathbf{y}$) if $\mathbf{x} \geqq \mathbf{y}$ but $\mathbf{x} \neq \mathbf{y}$ (i.e. the components of $\mathbf{x}$ are all greater than or equal to the corresponding components of $\mathbf{y}$, but at least one of them is actually greater).

The difference between these two concepts is quite obvious, and it is easy to visualize it geometrically in $\mathbb{R}^2$: positive vectors belong to the first quadrant excluding the axes, whereas non-negative ones belong to the first quadrant including the non-negative half-axes.

• *Sum.* Consider two vectors in $\mathbb{R}^n$, which, for instance, refer to the amounts of the same n types of goods in two warehouses

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}; \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

The total amount for the n types of goods in the two warehouses is plainly described by the vector

$$\begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_n + y_n \end{bmatrix}.$$

We indicate such a vector as $\mathbf{x} + \mathbf{y}$ and call it the *sum* of the two initial vectors. When the vectors under consideration refer to inventories, their components are necessarily non-negative numbers. On the other hand, if the involved vectors are, say, financial operations (i.e. incomes and outlays at the same maturities), the sum vector gathers the net cash flows at every maturity and its components might well be negative.

Given its definition, *the sum of two vectors in $\mathbb{R}^n$ is still a vector in $\mathbb{R}^n$*. We say that $\mathbb{R}^n$ is *closed* under addition. Moreover, the following properties hold:

a1. *commutativity:*

$$\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x} \quad \textit{for every } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n;$$

a2. *associativity:*

$$(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z}) \quad \textit{for every } \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n.$$

As a consequence, it is possible to write $\mathbf{x} + \mathbf{y} + \mathbf{z}$ without any brackets, because the "associations" of addends in groups do not modify the result.

a3. *There exists in $\mathbb{R}^n$ a vector, written $\mathbf{0}$, which, added to every vector $\mathbf{x}$, yields $\mathbf{x}$ itself. It is the zero vector, with all of its components equal to 0:*

$$\mathbf{x} + \mathbf{0} = \mathbf{0} + \mathbf{x} = \mathbf{x} \quad \textit{for every } \mathbf{x} \in \mathbb{R}^n.$$

a4. *Every vector $\mathbf{x}$ features an opposite, i.e. a vector $-\mathbf{x}$, which, added to $\mathbf{x}$, yields the zero vector. The vector $-\mathbf{x}$ can be obtained from $\mathbf{x}$ by switching the sign of every component:*

$$\mathbf{x} + (-\mathbf{x}) = (-\mathbf{x}) + \mathbf{x} = \mathbf{0} \quad \textit{for every } \mathbf{x} \in \mathbb{R}^n.$$

We leave to the reader to find some numerical examples to illustrate these four properties.

• *Product by a scalar.* Another operation is the multiplication of a vector by a real number, also called *scalar-vector product*[6]. It is performed by multiplying all of the components of a vector $\mathbf{x}$ by the same real number c; the result can be written in either of the following equivalent forms $c\mathbf{x}$ or $\mathbf{x}c$:

$$c\mathbf{x} = \mathbf{x}c = c \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} c := \begin{bmatrix} cx_1 \\ cx_2 \\ \vdots \\ cx_n \end{bmatrix}$$

For instance,

$$-4 \begin{bmatrix} 3 \\ -2 \\ 5 \end{bmatrix} = \begin{bmatrix} -12 \\ 8 \\ -20 \end{bmatrix}.$$

By this definition, if $\mathbf{x}$ is a vector with n components, $c\mathbf{x}$ is also a vector with n components: we say that $\mathbb{R}^n$ is *closed* under products by a scalar. Moreover, the following properties hold:

m1. *distributivity with respect to the sum between real numbers:*

$$(c + c')\,\mathbf{x} = c\mathbf{x} + c'\mathbf{x} \qquad \textit{for every } c, c' \in \mathbb{R} \textit{ and for every } \mathbf{x} \in \mathbb{R}^n;$$

m2. *distributivity with respect to the sum between vectors:*

$$c\,(\mathbf{x} + \mathbf{x}') = c\mathbf{x} + c\mathbf{x}' \qquad \textit{for every } c \in \mathbb{R} \textit{ and for every } \mathbf{x}, \mathbf{x}' \in \mathbb{R}^n;$$

m3. *associativity:*

$$c'\,(c\mathbf{x}) = (c'c)\,\mathbf{x} \qquad \textit{for every } c, c' \in \mathbb{R} \textit{ and for every } \mathbf{x} \in \mathbb{R}^n;$$

m4. *neutral element for scalar-vector product:*

$$1 \cdot \mathbf{x} = \mathbf{x} \qquad \textit{for every } \mathbf{x} \in \mathbb{R}^n.$$

From the properties listed above, the two following ones can be deduced.
If $\alpha \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$,

$$\begin{aligned} \alpha\mathbf{x} &= \mathbf{0} \quad \textit{if and only if either } \alpha = 0 \textit{ or } \mathbf{x} = \mathbf{0}; \\ (-1)\mathbf{x} &= -\mathbf{x} \qquad \textit{for every } \mathbf{x} \in \mathbb{R}^n. \end{aligned}$$

The figures represent the sum between vectors and the scalar-vector product in $\mathbb{R}^2$. The sum of two vectors $\mathbf{x}$ and $\mathbf{y}$ is obtained by the well-known *parallelogram rule*. The vector $\alpha\mathbf{x}$, for every α, belongs to the line passing through the origin and the point $\mathbf{x}$.

[6] The term *scalar* stands for "real number".

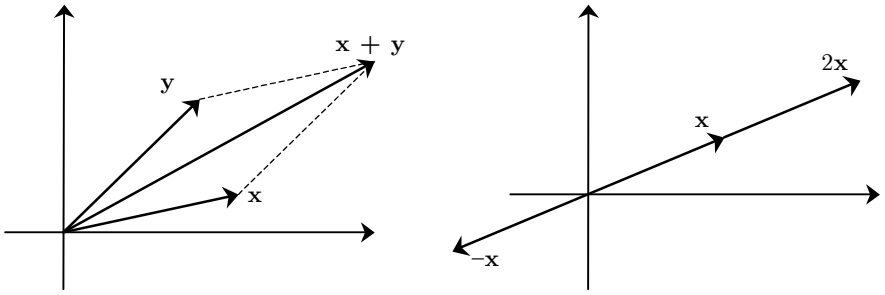


Figure 8.2. Sum of vectors and scalar-vector product

It is an interesting exercise to give a financial interpretation of these properties and to think of their intuitive meaning. Think of vectors as incomes and outlays at some given maturities, of sums of vectors as the cash flows generated by a portfolio of many such operations, of products by a scalar as a different scaling factor for a given operation represented by the vector $\mathbf{x}$. If, for instance, the vector

$$\mathbf{x} = \begin{bmatrix} -1000 \\ 100 \\ 1100 \end{bmatrix}$$

means that spending 1000 today (the first component is negative, so it stands for an outlay) we buy an asset which will pay 100 the next year and 1100 in two years, the vector $5\mathbf{x}$ gathers incomes and outlays generated by the purchase of 5 such assets, and $-4\mathbf{x}$ describes what happens when *issuing* four assets with the same characteristics. The reader can go on and interpret the four properties m1-m4 in a similar way.

• $\mathbb{R}^n$ *as a linear space.* The operations of sum and product by a scalar, with the properties a1-a4, m1-m4, provide $\mathbb{R}^n$ with the structure of a *linear space* (or *vector space*). Analogous operations, still called "sum " and "product by a scalar " (where the "scalars" might be either numbers of $\mathbb{R}$ or elements of another numerical field), can be defined on many other sets and still satisfy the $4 + 4$ properties seen above. Such sets, with this kind of structure, also become *linear spaces.*

8.2.1 Linear combinations

Suppose that k vectors are given, each describing a particular investment. If we enter each one of them with a certain exposure, it becomes interesting to evaluate the vector of the total cash flows of the portfolio. Such a vector is called a *linear combination* of the initial vectors. More precisely:

Definition 2.1. *Let* $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k$ *be vectors*[7] *of* $\mathbb{R}^n$ *and let* $c_1, c_2, \ldots, c_k$ *be real numbers. The vector (of* $\mathbb{R}^n$*)*

$$\mathbf{x} = \sum_{s=1}^{k} c_s \mathbf{x}^s = c_1\mathbf{x}^1 + c_2\mathbf{x}^2 + \ldots + c_k\mathbf{x}^k$$

is called a ***linear combination*** *of the vectors* $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k$ *with* ***weights*** *(or* ***coefficients****)* c_1, $c_2, \ldots, c_k$.

For instance, the vector $\begin{bmatrix} 4 \\ -10 \\ 8 \end{bmatrix}$ is a linear combination of the two vectors $\begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ -1 \\ -2 \end{bmatrix}$ with weights 4 and 2, respectively. Indeed,

$$4\begin{bmatrix} 1 \\ -2 \\ 3 \end{bmatrix} + 2\begin{bmatrix} 0 \\ -1 \\ -2 \end{bmatrix} = \begin{bmatrix} 4 \\ -10 \\ 8 \end{bmatrix}.$$

Every vector $\mathbf{x}$ in $\mathbb{R}^n$ is a linear combination of the fundamental vectors $\mathbf{e}^1, \ldots, \mathbf{e}^n$, and the coefficients of the linear combination turn out to be exactly the components of $\mathbf{x}$, i.e.

$$\mathbf{x} = \sum_{s=1}^{n} x_s \mathbf{e}^s.$$

Indeed:

$$\begin{aligned} \mathbf{x} &= \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ x_2 \\ \vdots \\ 0 \end{bmatrix} + \cdots + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ x_n \end{bmatrix} = \\ &= x_1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} + \cdots + x_n \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} = x_1\mathbf{e}^1 + x_2\mathbf{e}^2 + \cdots + x_n\mathbf{e}^n. \end{aligned}$$

- *Convex (linear) combinations.* When a linear combination

$$\sum_{s=1}^{k} c_s \mathbf{x}^s = c_1\mathbf{x}^1 + c_2\mathbf{x}^2 + \ldots + c_k\mathbf{x}^k$$

is such that $c_s \geq 0$ for all s and $c_1 + c_2 + \ldots + c_k = 1$, the linear combination is said to be *convex*.

[7] The reader should not confuse the superscripts used to index vectors with a power operation for vectors, which is not defined nor ever will be.

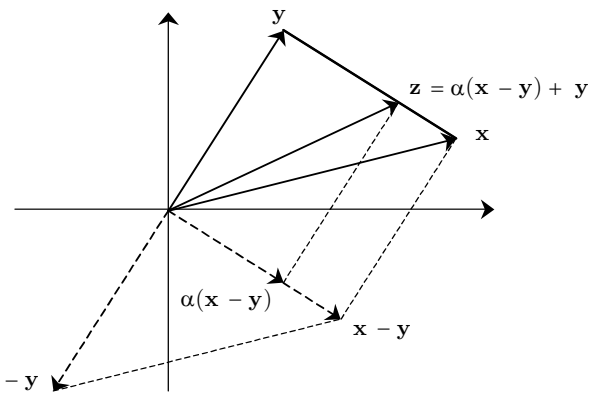


Figure 8.3. Convex combination of $\mathbf{x}$ and $\mathbf{y}$

The convex combinations of two vectors $\mathbf{x}$ e $\mathbf{y}$ in the cartesian plane (or in the space)

$$\mathbf{z} = \alpha\mathbf{x} + (1-\alpha)\,\mathbf{y} = \alpha\,(\mathbf{x}-\mathbf{y}) + \mathbf{y}, \qquad \alpha \in [0,1]\,,$$

describe the line *segment* connecting $\mathbf{x}$ and $\mathbf{y}$.

8.3 Inner product of two vectors

We now define another operation which is used in several applications and which does not depend on the "typographic style" (row or column) used to write two vectors. Such an operation is called the inner product or scalar product.

Definition 3.1. *Given two vectors* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, *the real number*

$$\mathbf{x} \cdot \mathbf{y} = \sum_{s=1}^{n} x_s y_s$$

is called the ***inner product*** (*or* ***scalar product***) *of* $\mathbf{x}$ *and* $\mathbf{y}$.

Thus if $\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ -2 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}$, we obtain

$$\mathbf{x} \cdot \mathbf{y} = 1 \cdot (-2) + 2 \cdot 0 + (-2) \cdot 1 = -4.$$

The inner product is sometimes also denoted by $\langle \mathbf{x}, \mathbf{y} \rangle$. When the first and the second vector are, respectively, a row and a column vector, the notation $\mathbf{xy}$ can be used as well. Thus, for instance, if both $\mathbf{x}$ and $\mathbf{y}$ are column vectors, we can write $\mathbf{x}^{\mathrm{T}}\mathbf{y}$ instead of $\mathbf{x} \cdot \mathbf{y}$.

Example 3.1. In a certain production process, goods of n different kinds are used and, to produce a unit of the product, x_1 units of the first kind, x_2 of the second

kind,..., x_n of the n-th kind are needed. The vector $\mathbf{x}$, with components $x_1, \ldots, x_n$, is called a *unit base slip*. Roughly speaking, $\mathbf{x}$ is the "list of ingredients" needed to make one unit of the product.

Suppose now that p_1 be the unit price of the first kind of goods, p_2 the unit price for the second,..., p_n the unit price for the n-th, and let us denote the vector of unit prices for the various kinds of goods by $\mathbf{p}$. The unit price of production is then given by

$$p_1x_1 + p_2x_2 + \cdots + p_nx_n = \sum_{s=1}^{n} p_s x_s = \mathbf{p} \cdot \mathbf{x}.$$

The inner product satisfies the following conditions ($\mathbf{x}, \mathbf{y}, \mathbf{y}' \in \mathbb{R}^n,\ h, k \in \mathbb{R}$):

p1. *commutativity:*

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{x}$$

p2. *distributivity with respect to addition between vectors:*

$$(\mathbf{x} + \mathbf{x}') \cdot \mathbf{y} = \mathbf{x} \cdot \mathbf{y} + \mathbf{x}' \cdot \mathbf{y}$$

p3. *homogeneity (with respect to both factors):*

$$h\mathbf{x} \cdot k\mathbf{y} = h\,(\mathbf{x} \cdot k\mathbf{y}) = hk\,(\mathbf{x} \cdot \mathbf{y})$$

p4. *non-negativity:*

$$\mathbf{x} \cdot \mathbf{x} \geq 0 \quad \textit{and} \quad \mathbf{x} \cdot \mathbf{x} = 0 \ \textit{ if and only if } \ \mathbf{x} = \mathbf{0}$$

A product between numbers is zero if and only if at least one of the two factors is zero. The inner product between vectors, on the contrary, can be zero even if the factors are both non-zero vectors. For instance, if

$$\mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

(neither of them is the zero vector), we obtain $\mathbf{x} \cdot \mathbf{y} = 1 + (-1) = 0$. A graphical representation of the two vectors shows that they are *orthogonal*. It is not hard to understand that this is also the case for three-dimensional vectors. One then extends such a notion to n-dimensional vectors.

Definition 3.2. *Two vectors* $\mathbf{x}, \mathbf{y}$ *in* $\mathbb{R}^n$ *are said to be* ***orthogonal*** *if*

$$\mathbf{x} \cdot \mathbf{y} = 0.$$

For instance, the fundamental vectors $\mathbf{e}^1, \mathbf{e}^2 \ldots \mathbf{e}^n$ are pairwise orthogonal in $\mathbb{R}^n$. Of course, the zero vector is orthogonal to any vector.

8.3.1 Modulus, distance

We know that vectors with one, two or three components can be represented by arrows on an axis, in the cartesian plane or in the three-dimensional space. There

exists a natural definition for the length of a vector $\mathbf{x}$: it simply coincides with the *distance* of the point $\mathbf{x}$ from the origin $\mathbf{0}$ and can be calculated (in two or three dimensions) by means of the Pythagorean Theorem. Such a length, also called the *modulus* or the *norm* of $\mathbf{x}$ and represented by $|\mathbf{x}|$, can be written as

$$|\mathbf{x}| = \begin{cases} \sqrt{x_1^2} = |x_1| & \text{in one dimension,} \\ \sqrt{x_1^2 + x_2^2} & \text{in two dimensions,} \\ \sqrt{x_1^2 + x_2^2 + x_3^2} & \text{in three dimensions.} \end{cases}$$

This concept can be extended to the general, n-dimensional case.

Definition 3.3. *The* ***modulus*** *or* (*Euclidean*[8]) ***norm*** *of a vector* $\mathbf{x} \in \mathbb{R}^n$ *is defined by the formula*

$$|\mathbf{x}| = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \sqrt{\sum_{s=1}^{n} x_s^2}.$$

For instance, the modulus of the vector

$$\mathbf{x} = \begin{bmatrix} 1 & -1 & 0 & 3 \end{bmatrix}^{\mathrm{T}}$$

is

$$|\mathbf{x}| = \sqrt{1^2 + (-1)^2 + 0^2 + 3^2} = \sqrt{11}.$$

The modulus satisfies the following conditions:

n1. *for every* $\mathbf{x} \in \mathbb{R}^n$, $|\mathbf{x}| \geq 0$ *and it is zero if and only if* $\mathbf{x} = \mathbf{0}$*;*

n2. $|c\mathbf{x}| = |c| \cdot |\mathbf{x}|$ *for every* $\mathbf{x} \in \mathbb{R}^n$ *and every scalar* c*;*

n3. $|\mathbf{x} + \mathbf{y}| \leq |\mathbf{x}| + |\mathbf{y}|$ *for every* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

The third property is called the *triangle inequality*, because in $\mathbb{R}^2$ or $\mathbb{R}^3$ it corresponds to a well-known theorem of elementary geometry about triangles: *the length of each side in a triangle cannot exceed the sum of the other two.*

We can also show that, for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, the following inequality, known as *Schwarz's*[9] *inequality,* holds:

$$\boxed{|\mathbf{x} \cdot \mathbf{y}| \leq |\mathbf{x}| \cdot |\mathbf{y}|} \tag{8.4}$$

Vectors with unit norm are called *unit vectors*. The fundamental vectors of $\mathbb{R}^n$ are unit vectors.

---

[8] Other norms can be defined, which are more useful or more natural to some extent. In some important parts of Statistics (e.g. classification methods), for instance, we might use the norm $\|\mathbf{x}\| = \sum_{s=1}^{n} |x_s|$, called the *taxi-driver norm* because, in the plane, its value is the distance from the origin and $\mathbf{x}$, as it would be measured by the fare meter of a taxi (honestly) travelling in a city laid out in rectangular blocks.

[9] Hermann Amandus Schwarz (1843-1921).

The norm allows us to define the distance between two vectors in $\mathbb{R}^n$. If we take into consideration two numbers x, y, or the two points representing them on a cartesian axis, their distance is $|x-y|$. We can proceed analogously with vectors in two or three dimensions. The usual distance notion in elementary geometry leads us to define

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(x_1-y_1)^2+(x_2-y_2)^2} \quad ; \quad d(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{s=1}^{3}(x_s-y_s)^2},$$

i.e. $d(\mathbf{x},\mathbf{y})$ is the norm of $\mathbf{x}-\mathbf{y}$. Geometrically, it is the length of the segment with extremal points $\mathbf{x}$ and $\mathbf{y}$.

The generalisation to n-dimensional vectors is straightforward.

Definition 3.4. *The (Euclidean)* ***distance*** $d(\mathbf{x},\mathbf{y})$ *between two vectors* $\mathbf{x},\mathbf{y}$ *in* $\mathbb{R}^n$ *is*

$$d(\mathbf{x},\mathbf{y}) = |\mathbf{x}-\mathbf{y}| = \sqrt{\sum_{s=1}^{n}(x_s-y_s)^2}.$$

The distance $d(\mathbf{x},\mathbf{y})$ satisfies the following conditions:

d1. *non negativity:* $d(\mathbf{x},\mathbf{y}) \geq 0$ *for every* $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ *and* $d(\mathbf{x},\mathbf{y}) = 0$ *if and only if* $\mathbf{x} = \mathbf{y}$;

d2. *symmetry:* $d(\mathbf{x},\mathbf{y}) = d(\mathbf{y},\mathbf{x})$ *for every* $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$;

d3. *the triangle inequality:*

$$d(\mathbf{x},\mathbf{y}) \leq d(\mathbf{x},\mathbf{z}) + d(\mathbf{z},\mathbf{y})$$

for every $\mathbf{x}$, $\mathbf{y}$ *and* $\mathbf{z}$ *in* $\mathbb{R}^n$.

For instance, given the vectors in $\mathbb{R}^4$

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 4 \\ 3 \\ 2 \\ 1 \end{bmatrix},$$

their distance is

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(1-4)^2+(2-3)^2+(3-2)^2+(4-1)^2} = 2\sqrt{5}.$$

8.4 Subspaces of $\mathbb{R}^n$

Given a vector $\mathbf{x} \neq \mathbf{0}$ in $\mathbb{R}^3$, let us consider the set of all vectors that can be written in the form $c\mathbf{x}$, with $c \in \mathbb{R}$. The geometrical representation of such a set is the line passing through $\mathbf{0}$ and $\mathbf{x}$. For instance, if

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

the vectors $c\mathbf{x}$ have components

$$\begin{cases} x_1 = c \\ x_2 = 2c \\ x_3 = 3c \end{cases} \tag{8.5}$$

and, as c varies throughout $\mathbb{R}$, it describes the line r passing through the origin and the point $\mathbf{x}$. Incidentally, equations (8.5) are called *parametric equations* of the straight line r.

Still in $\mathbb{R}^3$, let $\mathbf{x}^1$ and $\mathbf{x}^2$ be two non-zero vectors, which do not belong to the same line passing through the origin. Consider all of their linear combinations, i.e. the set composed of all of the vectors that can be written in the form $c_1\mathbf{x}^1+c_2\mathbf{x}^2$, with $c_1, c_2 \in \mathbb{R}$. The geometrical representation of such a set is the plane passing through $\mathbf{0}$, $\mathbf{x}^1$ and $\mathbf{x}^2$.

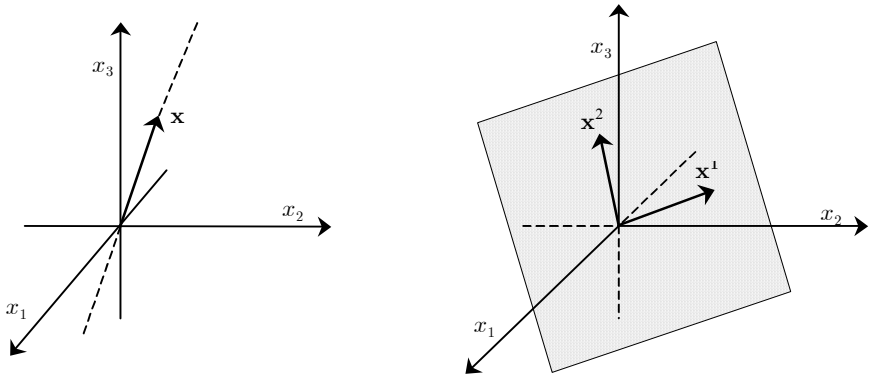


Figure 8.4. Line defined by a vector, plane defined by two vectors

For instance, if $\mathbf{x}^1 = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$ and $\mathbf{x}^2 = \begin{bmatrix} -1 \\ 1 \\ 3 \end{bmatrix}$, the vectors

$$\mathbf{x} = c_1 \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} + c_2 \begin{bmatrix} -1 \\ 1 \\ 3 \end{bmatrix}, \quad c_1, c_2 \in \mathbb{R} \tag{8.6}$$

identify the plane in $\mathbb{R}^3$ passing through the origin and the points $\mathbf{x}^1$ and $\mathbf{x}^2$. Writing (8.6) component by component, we obtain the *parametric equations* of the plane.

$$\begin{cases} x_1 = -c_2 \\ x_2 = c_1 + c_2 \\ x_3 = 2c_1 + 3c_2 \end{cases}$$

By solving the first equation with respect to c_2 and the second one with respect to c_1, and by substituting into the third equation, we obtain the *cartesian equation* of

the plane, given by:

$$x_3 = -x_1 + 2x_2.$$

Lines and planes of $\mathbb{R}^3$ passing through $\mathbf{0}$ are particular subsets of $\mathbb{R}^3$, the only ones which (besides the zero vector and $\mathbb{R}^3$ itself) satisfy the property of being *closed* under addition and multiplication by a scalar. For this reason they are called *linear subspaces.* To be precise:

Definition 4.1. *A non-empty subset*[10] V *in* $\mathbb{R}^n$ *such that, for every* $\mathbf{x}, \mathbf{y} \in V$ *and for every* $\alpha \in \mathbb{R}$,

$$\mathbf{x} + \mathbf{y} \in V, \qquad \alpha\mathbf{x} \in V,$$

is called a (*linear*) ***subspace*** *of* $\mathbb{R}^n$.

In other words, a subset V of $\mathbb{R}^n$ is a *subspace* if it is *closed* under the operation of linear combination, i.e. if

$$\alpha\mathbf{x} + \beta\mathbf{y} \in V \quad \text{for every } \mathbf{x}, \mathbf{y} \in V \text{ and every } \alpha, \beta \in \mathbb{R}.$$

Given k vectors $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k$ of $\mathbb{R}^n$, we denote by $\mathcal{C}(\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k)$ the set formed by *all* of their linear combinations. Thus, $\mathcal{C}(\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k)$ is the set of all of the vectors $\mathbf{x} \in \mathbb{R}^n$ which can be written in the form $\mathbf{x} = \sum_{s=1}^{k} c_s \mathbf{x}^s$.

Proposition 4.1. $\mathcal{C}(\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k)$ *is a subspace of* $\mathbb{R}^n$ *and is called the subspace spanned (or generated) by the vectors* $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k$.

Proof. Take two vectors in the set $\mathcal{C}(\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k)$, i.e. two vectors in the form

$$\mathbf{x} = \sum_{s=1}^{k} c_s \mathbf{x}^s \quad \text{and} \quad \mathbf{y} = \sum_{s=1}^{k} c'_s \mathbf{x}^s.$$

The vector $\alpha\mathbf{x} + \beta\mathbf{y}$ can be written as:

$$\alpha\mathbf{x} + \beta\mathbf{y} = \alpha \sum_{s=1}^{k} c_s \mathbf{x}^s + \beta \sum_{s=1}^{k} c'_s \mathbf{x}^s = \sum_{s=1}^{k} (\alpha c_s + \beta c'_s)\, \mathbf{x}^s,$$

which is still a linear combination of the k given vectors and thus belongs to $\mathcal{C}(\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k)$. □

We say that the vectors $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k$ constitute a *support* or a *spanning system* for the subspace. We can show that linearly combining a certain number of vectors of $\mathbb{R}^n$ is indeed the way to generate every possible subspace. In other words, for every subspace V of $\mathbb{R}^n$, it is possible to determine a set of vectors $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k$, such that $V = \mathcal{C}(\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k)$. As we shall see in the examples, every vector space has infinitely many spanning systems. Later on, we shall select some of them as being particularly significant.

[10] V may also coincide with $\mathbb{R}^n$ or with the zero vector alone.

Examples

4.1. Since for every bidimensional vector $\mathbf{x} \in \mathbb{R}^2$ we have

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

the pair of vectors $\mathbf{e}^1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $\mathbf{e}^2 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ is a support for $\mathbb{R}^2$.

Let us now consider $\mathbb{R}^3$. The three vectors

$$\mathbf{e}^1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}; \quad \mathbf{e}^2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}; \quad \mathbf{e}^3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

span $\mathbb{R}^3$, because:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{8.7}$$

Generally speaking, the fundamental vectors of $\mathbb{R}^n$ are a support for $\mathbb{R}^n$.

4.2. The set containing the zero vector (alone) is a particular subspace of $\mathbb{R}^n$. Notice that such a vector belongs to every subspace of $\mathbb{R}^n$.

4.3. The vectors

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \qquad \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

span $\mathbb{R}^2$. Indeed, every vector $\mathbf{x}$ can be written as a linear combination of the three given vectors, for instance like this:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = (x_1 - x_2) \begin{bmatrix} 1 \\ 0 \end{bmatrix} + x_2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + 0 \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix},$$

but also like this:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \left(x_1 - \frac{x_2}{2}\right) \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 0 \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \frac{x_2}{2} \cdot \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

4.4. Any three vectors of $\mathbb{R}^3$ span all of $\mathbb{R}^3$, if none of them is a linear combination of the other two. The vectors

$$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

span $\mathbb{R}^3$. Indeed, we can write

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + (x_2 - x_1) \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} + (x_3 - x_2) \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

The reader may identify other sets of vectors which span $\mathbb{R}^2$ and $\mathbb{R}^3$.

8.5 Linear dependence

Let us think of the vectors of $\mathbb{R}^n$ as financial operations which can be undertaken on various scales (both positive or negative). We shall call any linear combination of vectors with the meaning of financial operations a *portfolio*. We might wonder whether it is possible to reproduce some of these vectors by linearly combining other ones in a suitable portfolio. The question is particularly important from a financial point of view. If a given financial operation can be "artificially" reproduced (the term "synthetically" is often used), one might not take it explicitly into consideration, but rather think of it as a portfolio which can be obtained by other operations. In economical terms, this means that in a well organised market the price for such an operation is not "free", but strictly determined by the price of its "ingredients".

The concept of linear dependence that we introduce here summarizes the above lines.

Definition 5.1. *We say that k vectors $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k \in \mathbb{R}^n$ are **linearly dependent**, when at least one of them can be written as a linear combination of the others. If this is not the case, the vectors are called **linearly independent**.*

By possibly reordering the vectors, it is not restrictive to let the vector which can be written as a linear combination of the others be the last one, so that *linear dependence* means that there exist $k-1$ scalars $c_1, c_2, \ldots, c_{k-1}$ such that

$$\mathbf{x}^k = \sum_{s=1}^{k-1} c_s \mathbf{x}^s. \tag{8.8}$$

Example 5.1. The three vectors

$$\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

are linearly dependent. In particular, the third one is a linear combination of the other two. Let us show that indeed there exist $c_1, c_2 \in \mathbb{R}$, such that

$$c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}. \tag{8.9}$$

It is indeed sufficient to note that (8.9) is equivalent to

$$\begin{cases} c_1 = 2 \\ c_1 + c_2 = 1 \end{cases} \Rightarrow \begin{cases} c_1 = 2 \\ c_2 = -1 \end{cases}$$

and thus

$$\begin{bmatrix} 2 \\ 1 \end{bmatrix} = 2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

An equivalent definition of linear dependence/independence is expressed in the following proposition.

Proposition 5.1. *The vectors* $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k \in \mathbb{R}^n$ *are linearly dependent if and only if there exist* k *scalars* $c_1, \ldots, c_k$ *not all equal to zero, such that*

$$\sum_{s=1}^{k} c_s \mathbf{x}^s = \mathbf{0} \tag{8.10}$$

The vectors $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k \in \mathbb{R}^n$ *are linearly independent if and only if (8.10) holds only when all of the coefficients* $c_1, \ldots, c_k$ *are zero.*

Proof. Suppose that the vectors are linearly dependent, i.e. that (8.8) holds. By carrying $\mathbf{x}^k$ to the right hand side, we get

$$\mathbf{0} = \sum_{s=1}^{k-1} c_s \mathbf{x}^s - \mathbf{x}^k.$$

This way, the zero vector is a linear combination of the vectors under consideration and at least one of the coefficients is not zero, as $c_k = -1 \neq 0$.

On the contrary, suppose that (8.10) holds, for instance, with $c_k \neq 0$. Then we can write

$$\mathbf{0} = \sum_{s=1}^{k-1} c_s \mathbf{x}^s + c_k \mathbf{x}^k$$

whence

$$\mathbf{x}^k = -\frac{1}{c_k} \sum_{s=1}^{k-1} c_s \mathbf{x}^s = \sum_{s=1}^{k-1} \left(-\frac{c_s}{c_k} \right) \mathbf{x}^s,$$

which shows $\mathbf{x}^k$ as a linear combination of the other vectors. □

Example 5.2. The n fundamental vectors in $\mathbb{R}^n$ are linearly independent. Indeed, the equality

$$\sum_{s=1}^{n} c_s \mathbf{e}^s = \mathbf{0}$$

holds true only if every c_s is equal to 0.

Remarks

5.1. If a set of vectors contains the zero vector, they are necessarily linearly dependent.

5.2. Two non-zero vectors are linearly dependent if their components are proportional, i.e. if each of them is a "multiple" of the other.

5.3. We have seen that, given a set of linearly dependent vectors, it is possible to write one of them as a linear combination of the remaining ones. This does not mean that it is possible to write *any* one of them as a linear combination of the others. For instance, the vectors

$$\begin{bmatrix} 2 \\ -1 \end{bmatrix}, \qquad \begin{bmatrix} -6 \\ 3 \end{bmatrix}, \qquad \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

are linearly dependent, as

$$\begin{bmatrix} 2 \\ -1 \end{bmatrix} = -\frac{1}{3}\begin{bmatrix} -6 \\ 3 \end{bmatrix} + 0\begin{bmatrix} 3 \\ 1 \end{bmatrix}.$$

The first vector can be written as a linear combination of the other two, but nevertheless there is no linear combination of the first two which gives the third.

Moreover, as the reader can easily check, the following properties hold.

l1. *If we add some vectors to a set of linearly dependent vectors, linear dependence still holds. If we subtract some vectors from a set of linearly independent vectors, linear independence still holds.*

l2. *Given a set of linearly dependent vectors* $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k \in \mathbb{R}^n$, *if one of them is replaced by a multiple of itself linear dependence is preserved. The same happens for linear independence, with non-zero multiples.*

l3. *Given a set of linearly dependent vectors* $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k \in \mathbb{R}^n$, *when adding to one of them a linear combination of the others linear dependence is preserved. The same is true for the case of linear independence.*

l4. *If the vectors* $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k \in \mathbb{R}^n$ *are linearly independent, then the vectors* $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k, \mathbf{y} \in \mathbb{R}^n$ *are linearly dependent if and only if* $\mathbf{y}$ *can be written as a linear combination of* $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k$.

• (⇛**Chapter 11**) *A simple financial market.* Consider a financial market where assets are traded, expiring not later than after two years. Whoever underwrites such a contract pays immediately (at the date 0) a sum which gives the right to draw, 1 and/or 2 years later, some cash amounts fixed by contract. The following vectors describe three such assets. The components denote the cash movements at the dates $0, 1, 2$.

$$\begin{bmatrix} -100 \\ 110 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} -100 \\ 0 \\ 121 \end{bmatrix}, \quad \begin{bmatrix} -100 \\ 10 \\ 110 \end{bmatrix}.$$

The first two assets are *pure discount* ones, whereas the third one allows the withdrawal of interests at the end of each year. They are three different financial devices, each responding to a specific requirement:

— the first asset serves those willing to invest for exactly one year,

— the second one serves those willing to invest over two years, without the need for withdrawing interests during the process,

— the third one resembles the second, but allows for sums of interest to be obtained along the way and, among the three, this is the only one with such a feature.

We now show that the third asset is useless, meaning that it could be synthesized starting from the other two, suitably mixed in a portfolio. In fact, we have

$$\begin{bmatrix} -100 \\ 10 \\ 110 \end{bmatrix} = \alpha\begin{bmatrix} -100 \\ 110 \\ 0 \end{bmatrix} + \beta\begin{bmatrix} -100 \\ 0 \\ 121 \end{bmatrix}$$

where the coefficients α and β can be found by solving the system

$$\begin{cases} -100\alpha - 100\beta = -100 \\ 110\alpha = 10 \\ 121\beta = 110 \end{cases} \quad \Longrightarrow \quad \begin{cases} \alpha = 1/11 \\ \beta = 10/11. \end{cases}$$

To reproduce the third asset, then, it is sufficient to buy 1/11 units of the first one (which costs 100/11 and at maturity produces the income 10) and 10/11 units of the second (which costs 1000/11 and will yield 110 in two years' time).

Let us remark that, if we now wanted to get rid of another asset – either the first or the second one – we would not be able to reproduce it any longer. We also note that, instead of suppressing the third asset, we could have suppressed either of the other two, without reducing the opportunities offered by the market. This might look surprising, because the structure of the first two assets makes them natural "bricks" to build up the third, but how could one reproduce a brick with a complex object such as the third asset? The trick consists in using a negative amount of it. In a financial setting this is naturally possible, because buying a negative amount of an asset simply means *issuing* such an asset, or selling it without actually having it (such an operation is called "short selling").

One more remark: the cost of the "syntethic" assets matches the cost of the real assets perfectly. If this were not the case, some quite strange things would happen in the market. Suppose for a while that the third asset be slightly more expensive than the synthetic one, for instance that it involved an outlay of -101 instead of -100 to guarantee the corresponding inflows. In such a case, any financial operator realising it could make a profit by issuing such an asset and simultaneously buying a suitable mix of the first two assets which provide in the next 1 and 2 years the funds needed to honour the issued commitment. Playing such a game on a single unit of the third asset would yield an immediate income of 1 Euro, whereas on 10 contracts it would lead to 10 Euros and so on... It is indeed a *money pump*, i.e. a "tap" which produces money. Such an opportunity is more formally called a *riskless arbitrage*. In a financial market at equilibrium, arbitrages are not possible or, if they were, they would be available for a very short time, as the game we have just described would fatally lead to a decrease of the price of the third asset (massively supplied) and/or to an increase of the price of the first two (massively demanded), until the linear dependence between the three vectors were finally restored.

It is interesting to notice that all the three vectors are orthogonal to the vector

$$\begin{bmatrix} 1 \\ 1/1.1 \\ 1/1.21 \end{bmatrix}$$

which contains the discount factors at a 10% rate, i.e. the interest rate that we implicitly assumed to support the equilibrium of the given market. This points out another important characteristic of orthogonality between vectors, i.e. it can be seen as a characterising condition for the efficiency of a market. Indeed, when a market is working properly, the vectors describing financial operations are all orthogonal to the vector identified by the equilibrium interest rate (or, more generally, the several equilibrium interest rates at the various maturities).

8.6 Bases and dimension of a subspace of $\mathbb{R}^n$

Let us still interpret vectors as incomes and outlays generated by assets. Knowing that any asset listed on Piazza Affari in Milan can be "reproduced" starting from, say, a million assets is a fairly useless piece of information. On the contrary, it would be *very* interesting to know that a few assets are sufficient to span all of the others, because in such a case it would be sufficient to watch those assets in order to know the entire market. In vector language, the question is the following. What is the *minimum number* of vectors which allows us to span a subspace of $\mathbb{R}^n$? To answer the question, we now introduce the concept of a *basis* of a linear space V, a subspace of $\mathbb{R}^n$.

Definition 6.1. *Let $V \subseteq \mathbb{R}^n$ be a vector space. A set of vectors $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k$ in V, which are both linearly independent and a spanning system for V, is said to be a **basis** of V. The number of vectors in a basis of V is called the **dimension** of V and is denoted by* $\dim(V)$.

Definition 6.1 is sensible, as it is possible to prove that if V admits a basis of k vectors, then any other basis of V still features k vectors. It follows that the *number of vectors in a basis* is a property of the space itself, and not of the particular support we choose.

Examples

6.1. $\mathbb{R}^n$ has dimension n and the fundamental vectors are a basis for it.

6.2. The plane with equation $x_1 = 0$, i.e. the set of vectors with the first component equal to zero, is a subspace of $\mathbb{R}^3$ (every linear combination of vectors having the first component equal to zero still has the first component equal to zero). A basis of such a subspace is

$$\mathbf{e}^2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \qquad \mathbf{e}^3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

but any other pair of vectors with the first component equal to zero will do, as long as they are linearly independent. For instance, we could take the vectors

$$\mathbf{a}^1 = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}, \qquad \mathbf{a}^2 = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}.$$

It is easy to check that every vector $\mathbf{x}$ with the first component equal to zero is a linear combination of $\mathbf{a}^1$ and $\mathbf{a}^2$. In particular, we have

$$\begin{bmatrix} 0 \\ x_2 \\ x_3 \end{bmatrix} = x_2 \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} + (-x_3 + 2x_2) \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}.$$

It is also possible to prove that if $\dim(V) = k$, any choice of k linearly independent vectors in V always constitutes a basis for V. Moreover, the representation of the vectors of a subspace as a linear combination of the vectors in a basis is *unique*.

Theorem 6.1. *Let $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^k$ be a basis of the linear space V. Then every vector $\mathbf{y}$ can be written in the form*

$$\mathbf{y} = \sum_{s=1}^{k} c_s \mathbf{x}^s$$

and the coefficients c_s $(s = 1, \ldots, k)$ are uniquely identified.

We know that m vectors $\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^m$ in $\mathbb{R}^n$ span a subspace $V = (\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^m)$, whose dimension $\dim(V)$ is less than or equal to both m and n. How can we calculate $\dim(V)$?

In other words, given m vectors $\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^m$ in $\mathbb{R}^n$, how can we calculate how many among them are linearly independent?

Example 6.3. Consider the vectors

$$\mathbf{a}^1 = \begin{bmatrix} 2 \\ -1 \\ 0 \\ 3 \end{bmatrix}, \quad \mathbf{a}^2 = \begin{bmatrix} 1 \\ 3 \\ -2 \\ -4 \end{bmatrix}, \quad \mathbf{a}^3 = \begin{bmatrix} 3 \\ 2 \\ -2 \\ -1 \end{bmatrix}, \quad \mathbf{a}^4 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix},$$

and let us try to determine how many among them are linearly independent. The first two vectors are linearly independent (their components are not proportional). Let us check whether the first three vectors are independent. We easily find that the third one is a linear combination of the first two (it is their sum). Let us discard the third vector and see whether the fourth one is a linear combination of the first two, i.e. if it is possible to find two coefficients c_1 and c_2 such that

$$c_1 \begin{bmatrix} 2 \\ -1 \\ 0 \\ 3 \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ 3 \\ -2 \\ -4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

Then we must have

$$\begin{cases} 2c_1 + c_2 = 0 \\ -c_1 + 3c_2 = 0 \\ -2c_2 = 1 \\ 3c_1 - 4c_2 = 1. \end{cases}$$

From the first equation we obtain $c_2 = -1/2$ and then, from the second one, we deduce that $c_1 = -3/2$, whereas the fourth one yields $c_1 = -1/3$, a contradiction. So, out of the four given vectors, only three are linearly independent (but not any three: for instance, the first two together with the fourth one will work).

The vectors $\mathbf{a}^1, \mathbf{a}^2$ and $\mathbf{a}^4$ provide a basis of a subspace V of $\mathbb{R}^4$ with dimension 3. We could check that $\mathbf{a}^3$ is a linear combination of $\mathbf{a}^1, \mathbf{a}^2$ and $\mathbf{a}^4$ with uniquely determined coefficients. In particular, we have

$$\mathbf{a}^3 = 1 \cdot \mathbf{a}^1 + 1 \cdot \mathbf{a}^2 + 0 \cdot \mathbf{a}^4.$$

The process seen in the example can work if we have any number of vectors in any number of components. A calculation process (and algorithm) is needed so that, maybe with the help of a computer, this problem might be solved in its full generality. To this purpose we shall introduce, later on, the concept of the *rank* of a matrix.

8.7 Matrices

We often need to work not with single vectors, but with *blocks* of vectors. For instance, if n assets are traded in a market, with incomes and outlays spread over m maturities, we are dealing with n (column) vectors of $\mathbb{R}^m$:

$$\begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix}; \quad \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix}; \quad \cdots \quad ; \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix}$$

where the first index distinguishes the component of the vector and the second one denotes the vector itself. It is fully natural to deal with the array

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} = \left[\mathbf{a}^1\, \mathbf{a}^2 \dots \mathbf{a}^n\right].$$

Such a "table" is called a *matrix* with m rows and n columns, and can be denoted by $\mathbf{A}$ or $[a_{rs}]$ ($r = 1, 2, \dots, m$ and $s = 1, 2, \dots, n$). We say that $\mathbf{A}$ has *type* $m \times n$. By matching the vectors

$$\begin{bmatrix} 2 \\ -1 \\ 5 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 \\ 0 \\ 6 \end{bmatrix},$$

we can generate the matrix of type 3×2

$$\begin{bmatrix} 2 & 1 \\ -1 & 0 \\ 5 & 6 \end{bmatrix}. \tag{8.11}$$

In many ways, a matrix can be assimilated to a vector with mn components, arranged in an "accordion" pattern:

$$\begin{matrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{matrix} \quad \searrow\nearrow \quad \begin{matrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{matrix} \quad \searrow\nearrow \quad \cdots \quad \searrow\nearrow \quad \begin{matrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{matrix}$$

In this sense, the set of all matrices of a given type $m \times n$ is nothing else but $\mathbb{R}^{mn}$. For matrices, thus, we can introduce the concepts of equality, strong and weak inequality, and the related properties which we have already seen for vectors.

A matrix of type $m \times n$ can also by obtained by the matching of m row vectors, each of $\mathbb{R}^n$:

$$\begin{matrix}[a_{11} & a_{12} & \cdots & a_{1n}] \\ [a_{21} & a_{22} & \cdots & a_{2n}] \\ & & \vdots & \\ [a_{m1} & a_{m2} & \cdots & a_{mn}]\end{matrix} = \begin{bmatrix} \boldsymbol{\alpha}^1 \\ \boldsymbol{\alpha}^2 \\ \vdots \\ \boldsymbol{\alpha}^m \end{bmatrix}.$$

For instance, the matrix (8.11) can be thought of as obtained by the matching of the three row vectors $\begin{bmatrix} 2 & 1 \end{bmatrix}$, $\begin{bmatrix} -1 & 0 \end{bmatrix}$, $\begin{bmatrix} 5 & 6 \end{bmatrix}$, each in $\mathbb{R}^2$.

Vectors are particular examples of matrices: a column vector is a matrix of type $m \times 1$ (a single column), while a row vector is a $1 \times n$ matrix (a single row). It is sensible to distinguish between column and row vectors, when they are seen as matrices of these particular types.

• *Transposition.* Consider a matrix $\mathbf{A}$ of type $m \times n$. Exchange the rows and the columns. We get a matrix of type $n \times m$, called the *transpose matrix* of $\mathbf{A}$ and denoted by $\mathbf{A}^{\mathrm{T}}$. If

$$\mathbf{A} = \begin{bmatrix} 8 & 7 \\ -3 & 4 \\ 5 & 0 \end{bmatrix},$$

its transpose matrix is

$$\mathbf{A}^{\mathrm{T}} = \begin{bmatrix} 8 & -3 & 5 \\ 7 & 4 & 0 \end{bmatrix}.$$

Obviously, $\left(\mathbf{A}^{\mathrm{T}}\right)^{\mathrm{T}} = \mathbf{A}$.

• *Square matrices.* Matrices of type $n \times n$ are called *square matrices of order* n. The elements identified by equal indices, $a_{11}, a_{22}, ..., a_{nn}$, form the *main diagonal*.

A square matrix is called *symmetric* if it coincides with its transpose. In a symmetric matrix, elements which are symmetric with respect to the main diagonal are equal: $a_{ij} = a_{ji}$. A square matrix whose elements below (above) the main diagonal are all equal to zero is called an upper (a lower) *triangular* matrix.

The following square matrices of order three

$$\begin{bmatrix} 2 & -1 & 5 \\ -1 & 3 & 7 \\ 5 & 7 & -4 \end{bmatrix}, \quad \begin{bmatrix} 2 & -1 & 5 \\ 0 & 3 & 7 \\ 0 & 0 & -4 \end{bmatrix}, \quad \begin{bmatrix} 2 & 0 & 0 \\ -1 & 3 & 0 \\ 5 & 7 & -4 \end{bmatrix}$$

are respectively symmetric, upper triangular and lower triangular.

An order n square matrix of the form

$$\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{bmatrix},$$

that is, with any elements $\lambda_1, \lambda_2, \ldots, \lambda_n$ along the main diagonal and zero elements elsewhere, is called a *diagonal matrix*[11].

• *Submatrices.* By suppressing $m - p$ rows and $n - r$ columns in a matrix $\mathbf{A}$ of type $m \times n$, we obtain a *submatrix* of $\mathbf{A}$ of type $p \times r$. Let

$$\mathbf{A} = \begin{bmatrix} 1 & -2 & 3 & 6 & 4 \\ 5 & -5 & -4 & 2 & 1 \\ 0 & 3 & -5 & 3 & 0 \end{bmatrix}.$$

The two matrices

$$\mathbf{B} = \begin{bmatrix} 1 & -2 & 3 \\ 5 & -5 & -4 \\ 0 & 3 & -5 \end{bmatrix} \quad \text{and} \quad \mathbf{C} = \begin{bmatrix} -5 & 1 \\ 3 & 0 \end{bmatrix}$$

are submatrices of $\mathbf{A}$. $\mathbf{B}$ is obtained by suppressing the last two columns, $\mathbf{C}$ is obtained by suppressing the first row and the first, third and fourth columns.

8.8 Operations with matrices

8.8.1 Sum of matrices and product of a matrix by a scalar

The sum of matrices and the product of a matrix by a scalar are defined in the same way as for vectors. We introduce them by means of simple examples.

• *Sum.* We have seen that it is possible to add together only vectors with the same number of components. Analogously, only matrices of the same type (i.e. with the same number of rows and columns) can be added together. Consider

$$\mathbf{A} = \begin{bmatrix} 3 & 4 & -3 \\ -1 & 0 & 6 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 0 & -2 & 4 \\ 5 & -3 & -6 \end{bmatrix}.$$

Their sum is obtained by adding up the corresponding elements

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 3+0 & 4+(-2) & -3+4 \\ -1+5 & 0+(-3) & 6+(-6) \end{bmatrix} = \begin{bmatrix} 3 & 2 & 1 \\ 4 & -3 & 0 \end{bmatrix}.$$

The sum between matrices satisfies the same conditions a1-a4 as the sum between vectors. A *null* matrix, i.e. with elements all equal to 0, denoted by $\mathbf{O}$, is the "neutral element" with respect to the sum.

• *Product by a scalar.* Let

$$\mathbf{A} = \begin{bmatrix} 3 & 2 & -1 \\ -5 & 6 & 0 \end{bmatrix}.$$

[11] All (or only some) of the elements along the main diagonal might also be null! A diagonal matrix is obviously both upper and lower triangular.

The product $-3\mathbf{A}$ is

$$-3\mathbf{A} = \begin{bmatrix} -3\cdot 3 & -3\cdot 2 & -3\cdot(-1) \\ -3\cdot(-5) & -3\cdot 6 & -3\cdot 0 \end{bmatrix} = \begin{bmatrix} -9 & -6 & 3 \\ 15 & -18 & 0 \end{bmatrix}.$$

The multiplication of a matrix by a scalar satisfies the four conditions m1-m4. This means that the set $\mathcal{M}(m,n)$ of all matrices of a given type is a linear space.

8.8.2 Product of matrices

There are many ways the product between two matrices $\mathbf{A}$ and $\mathbf{B}$ might be defined. The one which is commonly used (the product "*rows by columns*") requires the number of columns in the first factor to be equal to the number of rows in the second one. In such a case we say that $\mathbf{A}$ and $\mathbf{B}$ are *conformable* or *composable*. Let $\mathbf{A}$ be a matrix of type $m \times n$, considered as a matching of row vectors, hence

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}^1 \\ \boldsymbol{\alpha}^2 \\ \vdots \\ \boldsymbol{\alpha}^m \end{bmatrix},$$

and $\mathbf{B}$ a matrix of type $n \times p$, considered as a matching of column vectors,

$$\mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1p} \\ b_{21} & b_{22} & \dots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \dots & b_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{b}^1 & \mathbf{b}^2 & \dots & \mathbf{b}^p \end{bmatrix}.$$

Their *product* $\mathbf{AB}$ is the matrix $\mathbf{C}$ with m rows (as many as $\mathbf{A}$) and p columns (as many as $\mathbf{B}$) whose generic element c_{ij} is obtained by performing the inner product between the row vector $\boldsymbol{\alpha}^i$ and the column vector $\mathbf{b}^j$:

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} \boldsymbol{\alpha}^1\mathbf{b}^1 & \boldsymbol{\alpha}^1\mathbf{b}^2 & \dots & \boldsymbol{\alpha}^1\mathbf{b}^p \\ \boldsymbol{\alpha}^2\mathbf{b}^1 & \boldsymbol{\alpha}^2\mathbf{b}^2 & \dots & \boldsymbol{\alpha}^2\mathbf{b}^p \\ \vdots & \vdots & & \vdots \\ \boldsymbol{\alpha}^m\mathbf{b}^1 & \boldsymbol{\alpha}^m\mathbf{b}^2 & \dots & \boldsymbol{\alpha}^m\mathbf{b}^p \end{bmatrix}.$$

More precisely: *Given two matrices $\mathbf{A} = [a_{ij}]$ and $\mathbf{B} = [b_{jk}]$ of type $m \times n$ and $n \times p$ respectively, the* ***product*** *of $\mathbf{A}$ and $\mathbf{B}$, denoted by $\mathbf{AB}$, is the matrix $\mathbf{C}$ of type $m \times p$ whose general element c_{ik} is determined by the following formula:*

$$c_{ik} = \boldsymbol{\alpha}^i\mathbf{b}^k = a_{i1}b_{1k} + a_{i2}b_{2k} + \cdots + a_{in}b_{nk} = \sum_{s=1}^{n} a_{is}b_{sk}.$$

If

$$\mathbf{A} = \begin{bmatrix} 1 & -2 \\ -3 & 4 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 6 & 2 & 3 \\ 4 & 5 & 1 \end{bmatrix}$$

we obtain

$$\mathbf{AB} = \begin{bmatrix} 1\cdot 6 - 2\cdot 4 & 1\cdot 2 - 2\cdot 5 & 1\cdot 3 - 2\cdot 1 \\ -3\cdot 6 + 4\cdot 4 & -3\cdot 2 + 4\cdot 5 & -3\cdot 3 + 4\cdot 1 \end{bmatrix} = \begin{bmatrix} -2 & -8 & 1 \\ -2 & 14 & -5 \end{bmatrix}.$$

In particular, the product between a matrix of type $1 \times n$ and a matrix of type $n \times 1$ coincides with the inner product between vectors.

The product of a matrix and a conformable null matrix is the null matrix:

$$\mathbf{AO} = \mathbf{O} \quad \text{and} \quad \mathbf{OA} = \mathbf{O}.$$

We note, furthermore, that a scalar factor in a product between matrices can be placed in any position without modifying the result:

$$c(\mathbf{AB}) = (c\mathbf{A})\,\mathbf{B} = \mathbf{A}\,(c\mathbf{B})$$

and thus the meaning of the expression $c\mathbf{AB}$ is univocally determined.

Let us now see two examples in an economical-financial setting, where the product between matrices is featured in a very natural way.

Example 8.1. A firm sells four types of goods in three different geographical areas and the following matrix $\mathbf{Q}$, of type 3×4, contains in any row the amounts sold (let us say, in tons) of the goods of each type, respectively, in the first, the second and the third area:

$$\mathbf{Q} = \begin{bmatrix} 100 & 125 & 40 & 10 \\ 200 & 210 & 65 & 12 \\ 150 & 170 & 42 & 6 \end{bmatrix}.$$

Let us now build a second matrix, which will be called $\mathbf{P}$, collecting in the columns, for each type of the goods, the unit sale price per ton without taxes and with 20% value added tax:

$$\mathbf{P} = \begin{bmatrix} 10 & 10\cdot 1.2 \\ 12 & 12\cdot 1.2 \\ 18 & 18\cdot 1.2 \\ 20 & 20\cdot 1.2 \end{bmatrix} = \begin{bmatrix} 10 & 12 \\ 12 & 14.4 \\ 18 & 21.6 \\ 20 & 24 \end{bmatrix}.$$

The product matrix $\mathbf{QP}$

$$\mathbf{QP} = \begin{bmatrix} 100 & 125 & 40 & 10 \\ 200 & 210 & 65 & 12 \\ 150 & 170 & 42 & 6 \end{bmatrix} \begin{bmatrix} 10 & 12 \\ 12 & 14.4 \\ 18 & 21.6 \\ 20 & 24 \end{bmatrix}$$

is a matrix with three rows, as many as the geographical areas, and two columns, as many as the price types, which summarizes the wholesale proceeds of the firm:

$$\begin{bmatrix} 100 & 125 & 40 & 10 \\ 200 & 210 & 65 & 12 \\ 150 & 170 & 42 & 6 \end{bmatrix} \begin{bmatrix} 10 & 12 \\ 12 & 14.4 \\ 18 & 21.6 \\ 20 & 24 \end{bmatrix} = \begin{bmatrix} 3420 & 4104 \\ 5930 & 7116 \\ 4416 & 5299.2 \end{bmatrix}.$$

• (⇛ **Chapter 11**) *Net present value.* Consider m financial operations generating n cash flows at n certain maturities $t_1, t_2, \ldots, t_n$. Let us place these cash flows in an orderly way in the matrix:

$$\mathbf{A} = \begin{bmatrix} \boldsymbol{\alpha}^1 \\ \boldsymbol{\alpha}^2 \\ \vdots \\ \boldsymbol{\alpha}^m \end{bmatrix} = \begin{matrix} [a_{11} & a_{12} & \cdots & a_{1n}] & \leftarrow \text{operation } 1 \\ [a_{21} & a_{22} & \cdots & a_{2n}] & \leftarrow \text{operation } 2 \\ & & \vdots & & \\ [a_{m1} & a_{m2} & \cdots & a_{mn}] & \leftarrow \text{operation } m \end{matrix}$$

where each row collects the cash flows of a certain operation. Let ϕ be a function with the meaning of a *compound discount factor*: $\phi(t)$ is the present value of 1 Euro maturing in the next t years. Let $b_1, b_2, \ldots, b_n$ be the values assumed by ϕ in correspondence with the maturities $t_1, t_2, \ldots, t_n$, i.e.

$$b_1 = \phi(t_1), \quad b_2 = \phi(t_2), \ldots, \quad b_n = \phi(t_n),$$

and denote the column vector with components $b_1, b_2, \ldots, b_n$ by $\mathbf{b}$. The product

$$\boldsymbol{\alpha}^1\mathbf{b} = a_{11}b_1 + a_{12}b_2 + \cdots + a_{1n}b_n = \sum_{s=1}^{n} a_{1s}b_s = \sum_{s=1}^{n} a_{1s}\phi(t_s)$$

has the meaning of *net present value*[12] of the financial operation, calculated with respect to the discount law described by the function ϕ. The product matrix (column vector)

$$\mathbf{Ab} = \begin{bmatrix} \boldsymbol{\alpha}^1 \\ \boldsymbol{\alpha}^2 \\ \vdots \\ \boldsymbol{\alpha}^m \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}^1\mathbf{b} \\ \boldsymbol{\alpha}^2\mathbf{b} \\ \vdots \\ \boldsymbol{\alpha}^m\mathbf{b} \end{bmatrix}$$

contains the net present value for *all* of the operations gathered in the matrix $\mathbf{A}$.

For instance, consider two assets which generate, at the maturities $0, 1, 2$, the cash flows described by the rows of the matrix

$$\begin{bmatrix} -1000 & 100 & 1100 \\ -1000 & 0 & 1200 \end{bmatrix}$$

and let

$$\begin{bmatrix} 1 \\ 1/1.1 \\ 1/(1.1)^2 \end{bmatrix}$$

be the column vector (in $\mathbb{R}^3$) of the compound discount factors with interest rate 10% in correspondence with the maturities $0, 1, 2$. The vector which contains the net

[12] It is called "net" to stress the fact that the present value of outlays is taken away from the present value of incomes.

present values of the two investments is

$$\begin{bmatrix} -1000 & 100 & 1100 \\ -1000 & 0 & 1200 \end{bmatrix} \begin{bmatrix} 1 \\ 1/1.1 \\ 1/(1.1)^2 \end{bmatrix} = \begin{bmatrix} 0 \\ -8.26 \end{bmatrix}.$$

Properties of the product

Among the many properties of the product between matrices[13] we point out the following ones.

M1. *Associativity:*

$$(\mathbf{AB})\,\mathbf{C} = \mathbf{A}\,(\mathbf{BC}),$$

which legitimates the omission of brackets in the product $\mathbf{ABC}$.

M2. *Right and left distributive property with respect to addition:*

$$\mathbf{A}\,(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC} \quad \text{and} \quad (\mathbf{B}+\mathbf{C})\,\mathbf{A} = \mathbf{BA}+\mathbf{CA}.$$

M3. *Product and transposition:*

$$(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}.$$

Warning! The product between matrices *is not commutative*. If we modify the ordering between the factors in a product, it could become no longer defined or, even if it were, it might yield a different result. A 2×3 matrix can be multiplied by a 3×5 one and the result is a 2×5 matrix. The inverse product is not even defined. Even when the two products exist, they are generally different. For instance,

$$\begin{bmatrix} 2 & 3 \\ 4 & -1 \end{bmatrix} \begin{bmatrix} 5 & 0 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 13 & 6 \\ 19 & -2 \end{bmatrix}$$

whereas

$$\begin{bmatrix} 5 & 0 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 4 & -1 \end{bmatrix} = \begin{bmatrix} 10 & 15 \\ 10 & 1 \end{bmatrix}.$$

Therefore, when a matrix $\mathbf{A}$ is multiplied by a matrix $\mathbf{B}$, it is generally necessary to specify whether it is *premultiplied* ($\mathbf{BA}$) or *postmultiplied* ($\mathbf{AB}$). Of course, in some cases it may happen that $\mathbf{AB} = \mathbf{BA}$. In such a case, we says that the matrices $\mathbf{A}$ and $\mathbf{B}$ *commutate*.

Another property of multiplication between scalars which no longer holds for matrices is the zero product property. A product between numbers is zero *if* and *only if* at least one of the factors is zero. We have indeed seen that $\mathbf{AO} = \mathbf{OA} = \mathbf{O}$, but it is not generally true that if $\mathbf{AB} = \mathbf{O}$ then either $\mathbf{A}$ or $\mathbf{B}$ is the null matrix. For instance,

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

[13] The reader is invited to check them.

If the product of two matrices is the null matrix, it simply means that the row vectors of the first matrix are all orthogonal to the column vectors of the second.

• *Matrix product and linear combinations.* The following remark will be useful later on. Consider the matrix $\mathbf{A} = \left[\mathbf{a}^1\ \mathbf{a}^2 \dots \mathbf{a}^n\right]$, where the columns have been highlighted. The product of such a matrix with the column vector $\mathbf{b} = [b_1\ b_2 \cdots b_n]^T$ is the vector

$$\begin{aligned}
\mathbf{Ab} &= \begin{bmatrix} a_{11}b_1 + a_{12}b_2 + \cdots + a_{1n}b_n \\ a_{21}b_1 + a_{22}b_2 + \cdots + a_{2n}b_n \\ \vdots \\ a_{m1}b_1 + a_{m2}b_2 + \cdots + a_{mn}b_n \end{bmatrix} = \\
&= b_1 \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix} + b_2 \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix} + \cdots + b_n \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{mn} \end{bmatrix} = \\
&= b_1\mathbf{a}^1 + b_2\mathbf{a}^2 + \ldots + b_n\mathbf{a}^n = \sum_{s=1}^{n} b_s\mathbf{a}^s.
\end{aligned}$$

Thus, the vector $\mathbf{Ab}$ turns out to be the linear combination of the columns of $\mathbf{A}$, where the coefficients in the combinations are the components of $\mathbf{b}$.

Diagonal, scalar and unit matrices.

By *postmultiplying* a generic matrix $\mathbf{A} = \left[\mathbf{a}^1\ \mathbf{a}^2 \cdots \mathbf{a}^n\right]$ of type $m \times n$ by a *diagonal* matrix $\mathbf{\Lambda}$ of order n (with elements $\lambda_1, \lambda_2, \ldots, \lambda_n$ in the main diagonal) we get

$$\mathbf{A\Lambda} = \left[\lambda_1\mathbf{a}^1\ \lambda_2\mathbf{a}^2 \cdots \lambda_n\mathbf{a}^n\right],$$

that is, the elements in the main diagonal of $\mathbf{\Lambda}$ multiply the corresponding columns of $\mathbf{A}$. By *premultiplying* $\mathbf{A}$ of type $m \times n$ with $\mathbf{\Lambda}$, diagonal matrix of order m, we would get the matrix

$$\mathbf{\Lambda A} = \begin{bmatrix} \lambda_1\boldsymbol{\alpha}^1 \\ \lambda_2\boldsymbol{\alpha}^2 \\ \vdots \\ \lambda_m\boldsymbol{\alpha}^m \end{bmatrix},$$

where the rows of $\mathbf{A}$ are multiplied by the elements of the main diagonal of $\mathbf{\Lambda}$.

Example 8.2. Consider a matrix $\mathbf{Q}$ of amounts of commodities sold by a firm (three types of goods and two geographical areas)

$$\mathbf{Q} = \begin{bmatrix} 100 & 125 & 80 \\ 80 & 95 & 70 \end{bmatrix}.$$

Let $\mathbf{P}$ be the diagonal matrix, which contains the unit sale price for each of the three types of goods along the main diagonal:

$$\mathbf{P} = \begin{bmatrix} 12 & 0 & 0 \\ 0 & 11 & 0 \\ 0 & 0 & 15 \end{bmatrix}.$$

The product matrix $\mathbf{QP}$:

$$\mathbf{QP} = \begin{bmatrix} 100 & 125 & 80 \\ 80 & 95 & 70 \end{bmatrix} \begin{bmatrix} 12 & 0 & 0 \\ 0 & 11 & 0 \\ 0 & 0 & 15 \end{bmatrix} = \begin{bmatrix} 1200 & 1375 & 1200 \\ 960 & 1045 & 1050 \end{bmatrix}.$$

contains the wholesale profit obtained for every region and every type of goods.

The diagonal matrices whose elements on the main diagonal are all equal: $\lambda_1 = \lambda_2 = \cdots = \lambda_n = \lambda$ are called *scalar matrices*. Pre- or postmultiplying any matrix $\mathbf{A}$ by a scalar matrix featuring the number λ on the main diagonal, we obtain the same product of the scalar λ by the matrix $\mathbf{A}$. For instance,

$$\mathbf{\Lambda A} = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} \lambda a & \lambda b \\ \lambda c & \lambda d \end{bmatrix} = \lambda \mathbf{A}.$$

An order n scalar matrix, then, commutates with any square matrix of the same order. The scalar matrix of order n, having the value 1 along the main diagonal is called the *unit matrix* of order n, and is written as $\mathbf{I}_n$:

$$\mathbf{I}_n = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}.$$

The matrix $\mathbf{I}_n$ is then obtained by matching (row by row, or column by column) the fundamental vectors of $\mathbb{R}^n$. Pre- or postmultiplying $\mathbf{I}_n$ (on the right or on the left) by a (conformable) matrix $\mathbf{A}$ we obtain $\mathbf{A}$. $\mathbf{I}_n$ is thus the neutral element with respect to the matrix product. For every square matrix $\mathbf{A}$ with order n, we have

$$\mathbf{I}_n\mathbf{A} = \mathbf{A}\mathbf{I}_n = \mathbf{A}.$$

A scalar matrix can be written as $\lambda\mathbf{I}_n$.

• *Power matrix of a matrix.* We can define the positive integer powers of a square matrix $\mathbf{A}$ of order n by setting

$$\mathbf{A}^1 = \mathbf{A}; \quad \mathbf{A}^2 = \mathbf{A} \cdot \mathbf{A}; \quad \dots \quad \mathbf{A}^k = \underbrace{\mathbf{A} \cdot \mathbf{A} \cdot \dots \cdot \mathbf{A}}_{k \text{ times}}$$

We could also set

$$\mathbf{A}^0 = \mathbf{I}_n,$$

as we do with powers of numbers. For instance, if

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ -3 & -2 & -1 \\ 0 & 1 & 0 \end{bmatrix}$$

we have:

$$\mathbf{A}^2 = \mathbf{A} \cdot \mathbf{A} = \begin{bmatrix} -5 & 1 & 1 \\ 3 & -3 & -7 \\ -3 & -2 & -1 \end{bmatrix}.$$

8.8.3 Inverse matrix

We have defined the product between matrices and the power matrices of a matrix. We wonder: is it also possible to define the *inverse matrix* of a given matrix? Let us try to quote the definition of the inverse of a real number. A number b is the inverse of a number $a\,(\neq 0)$, if the product $ab = ba$ equals 1. Now, the products $\mathbf{AB}$ and $\mathbf{BA}$ do not generally both exist at the same time. Moreover, even if they exist, they might be different. Let us then restrict ourselves, as we have done for the power matrices, to the case of *square* matrices with *the same order*. We have seen that for such matrices the role of the number 1 is played by the *unit* matrix. We have thus naturally come to the

Definition 8.1. *Given a square matrix* $\mathbf{A}$ *of order* n*, we say that* $\mathbf{B}$ *is the* ***inverse*** *of* $\mathbf{A}$ *if* $\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$.

It is possible to show that *if* $\mathbf{B}$ and $\mathbf{C}$ are both inverses of the same matrix $\mathbf{A}$, then $\mathbf{B} = \mathbf{C}$; in other words *the inverse matrix, if it exists, is unique*. Indeed, $\mathbf{B}$ and $\mathbf{C}$ both being inverses of $\mathbf{A}$,

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n \quad \text{and} \quad \mathbf{AC} = \mathbf{CA} = \mathbf{I}_n,$$

and we thus obtain

$$\mathbf{B} = \mathbf{BI}_n = \mathbf{B}\,(\mathbf{AC}) = (\mathbf{BA})\,\mathbf{C} = \mathbf{I}_n\mathbf{C} = \mathbf{C}.$$

If *the inverse* of the matrix $\mathbf{A}$ exists, $\mathbf{A}$ is called an *invertible* matrix and its inverse matrix is denoted by the symbol $\mathbf{A}^{-1}$.

Definition 8.2. *A matrix which is not invertible is called* ***singular***.

We now wonder which (square) matrices are invertible. The analogy with numbers does not work, as a number is invertible if it is different from zero. In the case of matrices, the question is more subtle. For instance, the null matrix is not invertible, but

$$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 2 & 0 \end{bmatrix}$$

which is not zero, is not invertible either. If it were invertible, indeed, there would exist a matrix $\mathbf{B} = \begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix}$ such that

$$\begin{bmatrix} \alpha & \beta \\ \gamma & \delta \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 2 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

By performing the rows by columns product, we obtain

$$\begin{bmatrix} \alpha + 2\beta & 0 \\ \gamma + 2\delta & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and such an equality is impossible, whatever the values $\alpha, \beta, \gamma, \delta$ are.

Two problems are left open:
— How can we estabilish if a matrix is invertible?
— How can we calculate the inverse matrix?
The answer to both questions can be based upon the concept of *determinant*, which is developed in the next section.

8.9 The determinant

Instead of the questions we considered at the end of the last section, let us slightly shift our point of view and ask ourselves:

Given n vectors in $\mathbb{R}^n$, how can one understand whether they are linearly dependent or independent?

Let us start by considering vectors in $\mathbb{R}^2$. Two vectors $[a \quad b]$ and $[c \quad d]$ are linearly dependent if their components are proportional, i.e. if it is possible to find a real number k such that

$$a = kc \quad \text{and} \quad b = kd. \tag{8.12}$$

Such a pair of equalities, on the other hand, is equivalent[14] to

$$ad - bc = 0. \tag{8.13}$$

The number $ad - bc$ is then important, and is called the *determinant* of the matrix

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{8.14}$$

obtained as a matching (row by row) of the two vectors $[a \quad b]$ and $[c \quad d]$. Summarizing, two vectors $[a \quad b]$ and $[c \quad d]$ are linearly *dependent if and only if the*

[14] Indeed, by multiplying the first equality in (8.12) by d and the second one by b we find that $ad = bc$. On the other hand, if $ad - bc = 0$ and $cd \neq 0$, then $a/c = b/d = k$, which leads to (8.12); if $c = d = 0$, then (8.12) holds with $k = 0$; if $c = 0$, $d \neq 0$, we deduce $a = 0$ and (8.12) holds with $k = b/d$. The other case is similar.

determinant of the matrix (8.14) *is zero.* We denote the determinant of a matrix $\mathbf{A}$ by one of the symbols

$$\det \mathbf{A}, \qquad |\mathbf{A}|$$

or, if we want to highlight the elements of the matrix, we use the notation $\begin{vmatrix} a & b \\ c & d \end{vmatrix}$.

We note that the determinant of the transpose matrix $\mathbf{A}^{\mathrm{T}}$ is still equal to $ad - bc$. It then makes no difference whether we think in terms of row or column vectors.

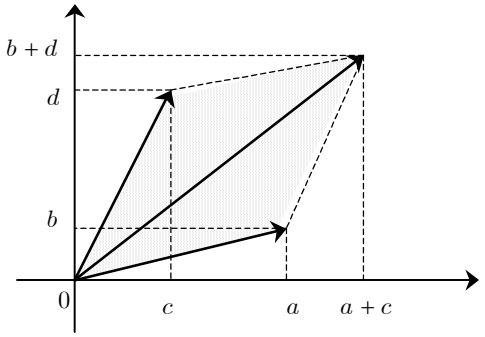


Figure 8.5. Determinant as area (with sign)

The absolute value of the determinant, $|ad - bc|$, has an interesting geometrical meaning. It equals the area of the parallelogram determined by the vectors $[a \quad b]$ and $[c \quad d]$. Its vertices are at the points $(0,0)$, (a,b), (c,d) and $(a+c, b+d)$, and its area is the product of the base length $\sqrt{a^2+b^2}$ by the height $|ad-bc| / \sqrt{a^2+b^2}$, which is the distance between the point (c,d) and the line passing through $(0,0)$ and (a,b), whose equation is $ay - bx = 0$. When the two vectors $[a \quad b]$ and $[c \quad d]$ are linearly dependent, the parallelogram reduces to a segment, whose area is of course zero: thus, $|ad - bc|$ keeps the meaning of an area in this case as well.

Inspired by the example in $\mathbb{R}^2$, we now define the determinant of a square matrix of order 3. It is possible to show, through calculations which are less simple than the ones we have just seen, that if

$$\mathbf{A} = \left[\mathbf{a}^1 \; \mathbf{a}^2 \; \mathbf{a}^3\right] = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

and we set $\det \mathbf{A}$ equal to the value

$$a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}, \qquad (8.15)$$

then $|\det \mathbf{A}|$ coincides with the volume of the parallelogram determined by the vectors $\mathbf{a}^1$, $\mathbf{a}^2$, $\mathbf{a}^3$, and it is equal to zero if and only if the three vectors are *linearly dependent*.

The expression for $\det \mathbf{A}$ can be easily kept in mind by recalling the following scheme (first suggested by the mathematician Sarrus).

Example 9.1. Let us calculate

$$\begin{vmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \\ -1 & -1 & 4 \end{vmatrix}.$$

First, we copy the first two columns on the right

$$\begin{array}{ccccccccc} 1 & & 2 & & 3 & & 1 & & 2 \\ & \searrow & & \swarrow\!\!\searrow & & \swarrow\!\!\searrow & & \swarrow & \\ 3 & & 2 & & 1 & & 3 & & 2 \\ & \swarrow & & \swarrow\!\!\searrow & & \swarrow\!\!\searrow & & \searrow & \\ -1 & & -1 & & 4 & & -1 & & -1 \end{array}$$

and then we calculate the products of the elements along the diagonals: adding the ones on the main diagonal ($1 \times 2 \times 4$) and on the "parallel" lines ($2 \times 1 \times (-1)$ and $(3 \times 3 \times (-1))$ - and subtracting those on the other original diagonal ($3 \times 2 \times (-1)$) and on the parallel lines ($1 \times 1 \times (-1)$ and $2 \times 3 \times 4$)):

$$8 - 2 - 9 - (-6 - 1 + 24) = -20.$$

The determinant (8.15) can also be written, by partially collecting the elements in the first row, as

$$a_{11}\left(a_{22}a_{33} - a_{23}a_{32}\right) + a_{12}\left(-a_{21}a_{33} + a_{23}a_{31}\right) + a_{13}\left(a_{21}a_{32} - a_{22}a_{31}\right).$$

The expressions in brackets are the determinants of the matrices obtained by cancelling the row and the column where the collected element belongs. The determinant multiplied by a_{12} gets, furthermore, the opposite sign. We thus obtain the formula

$$\det \mathbf{A} = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}, \tag{8.16}$$

which shows how the calculation of the determinant of an order 3 matrix can be led back to the calculation of the determinants of some order 2 matrices. Similar formulae can be obtained by collecting, instead of the elements in the first row, the elements of any other row or any column. Such formulae can be generalized to get an algorithm for the computation of the determinant of matrices of order greater than 3. We might also consider the case $n = 1$, by setting

$$\det\left[a_{11}\right] = a_{11}.$$

To generalize (8.16), some auxiliary concepts are needed. Let $\mathbf{B}$ be any matrix of type $m \times n$. If $k \leq \min(m, n)$, the determinant of any square submatrix of $\mathbf{B}$ of order k is called a **minor** of order k extracted from $\mathbf{B}$. The following definition is important.

Definition 9.1. *Let $\mathbf{A} = [a_{rs}]$ be a square matrix. The **complementary minor** of the element a_{rs} is the determinant M_{rs} of the submatrix obtained by suppressing the row and the column where such an element belongs. The number*

$$A_{rs} = (-1)^{r+s} M_{rs}$$

is called the ***algebraic complement*** (*or the* ***cofactor***) *of* a_{rs}.

We note that the algebraic complement of a_{rs} is equal to the complementary minor of a_{rs} when $r + s$ is even, and to its opposite when $r + s$ is odd.[15]. Let us calculate, for instance, the minor M_{21} and the algebraic complement A_{21} in the matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

We need to erase the second row and the first column, thus obtaining the submatrix

$$\begin{bmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{bmatrix}.$$

We then have:

$$M_{21} = \det \begin{bmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{bmatrix} = a_{12}a_{33} - a_{13}a_{32}, \qquad A_{21} = (-1)^{2+1} M_{21} = -M_{21}.$$

The algebraic complements of a row (or of a column) *do not* depend on the elements of the row (column) itself, i.e. they are *not* a function of the corresponding elements, but only depend on the place occupied by such elements.

Definition 9.2. *The* ***determinant*** *of a square matrix* $\mathbf{A}$ *is the sum of the products of the elements in the first column by their own algebraic complements:*

$$\det \mathbf{A} = \sum_{r=1}^{n} a_{r1} A_{r1}.$$

It is possible to prove that the same result is also obtained when performing the sum of the products of the elements in the first row by their own algebraic complements, i.e. that

$$\det \mathbf{A} = \sum_{s=1}^{n} a_{1s} A_{1s}.$$

[15] In a matrix, the positions where the sum of the indices is even are called *even places*, while the other ones are called *odd places*. By labelling the even places with a + and the odd ones with a −, we note that they assume a "chequered" layout:

$$\begin{bmatrix} + & - & + & \dots \\ - & + & - & \dots \\ + & - & + & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

More than that, *Laplace's theorem*[16] ensures that we can choose any line to calculate the determinant. In detail, in the case of the r-th row:

$$\det \mathbf{A} = \sum_{s=1}^{n} a_{rs} A_{rs},$$

and in the case of the s-th column:

$$\det \mathbf{A} = \sum_{r=1}^{n} a_{rs} A_{rs}.$$

Example 9.2. Let us calculate the determinant of the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 3 & -1 & 0 \\ -1 & 2 & 1 & 4 \\ 0 & 0 & -3 & 5 \\ 3 & 1 & 0 & 1 \end{bmatrix}.$$

To expand the determinant, it is convenient to choose a line which features the highest possible number of zero elements. For instance, the third row is fine. We thus obtain the following expression for $\det \mathbf{A}$:

$$\det \mathbf{A} = -3 \begin{vmatrix} 2 & 3 & 0 \\ -1 & 2 & 4 \\ 3 & 1 & 1 \end{vmatrix} - 5 \begin{vmatrix} 2 & 3 & -1 \\ -1 & 2 & 1 \\ 3 & 1 & 0 \end{vmatrix} =$$

and, by a further expansion, choosing the first row for the first determinant and the third row for the second one:

$$\begin{aligned} &= -3 \left(2 \begin{vmatrix} 2 & 4 \\ 1 & 1 \end{vmatrix} - 3 \begin{vmatrix} -1 & 4 \\ 3 & 1 \end{vmatrix} \right) - 5 \left(3 \begin{vmatrix} 3 & -1 \\ 2 & 1 \end{vmatrix} - \begin{vmatrix} 2 & -1 \\ -1 & 1 \end{vmatrix} \right) = \\ &= -3 \cdot 35 - 5 \cdot 14 = -175. \end{aligned}$$

8.9.1 Properties of the determinant

We now list the main properties of the determinant.

(*a*) *A matrix and its transpose have the same determinant*:

$$\det \mathbf{A} = \det \mathbf{A}^{\mathrm{T}}.$$

(*b*) *The determinant of a triangular matrix, in particular a diagonal one, is the product of the elements in the main diagonal.*

$$\det \mathbf{A} = a_{11} a_{22} \cdots a_{nn}.$$

[16] Pierre Simon, marquis de Laplace (1749-1827), besides being a distinguished mathematician, was also the domestic affairs minister for Napoleon, though he was removed from his office after six weeks.

If the matrix is *scalar*: $\mathbf{A} = a\mathbf{I}_n$ then $\det \mathbf{A} = a^n$. In particular, we have

$$\det \mathbf{I}_n = 1.$$

Thanks to property (a), almost all of the properties of the determinant can be stated indifferently by referring either to rows or to columns. So, in what follows, we shall adopt the term "line" to mean either a row or a column.

(c) *If a line is multiplied by a constant k the determinant gets multiplied by the same k.* For instance:

$$\begin{vmatrix} ka & kb \\ c & d \end{vmatrix} = kad - kbc = k\,(ad - bc) = k \begin{vmatrix} a & b \\ c & d \end{vmatrix}.$$

In particular (by setting $k = 0$): *if a line is zero, the determinant is* 0.

(d) *If two parallel lines are swapped, the determinant changes its sign.* For instance:

$$\begin{vmatrix} c & d \\ a & b \end{vmatrix} = cb - ad = -\,(ad - bc) = -\begin{vmatrix} a & b \\ c & d \end{vmatrix}.$$

(e) *If two parallel lines are equal, the determinant is* 0. Let D be the value of the determinant. By swapping the two equal lines, the determinant must become $-D$ by the preceding property. However, since we swapped two equal lines, we must have $D = -D$, whence $D = 0$.

(f) *If a line is a multiple of a parallel line, then the determinant is 0.* Indeed, if the line is k times a parallel line, we can "extract" k (property (c)) and revert to a matrix with two equal lines, whose determinant is 0 (property (e)).

(g) *If a line* $\mathbf{a}$ *is the sum of two vectors* $\mathbf{a} = \mathbf{a}' + \mathbf{a}''$*, the determinant is the sum of the determinants of the matrices obtained by replacing* $\mathbf{a}$ *with* $\mathbf{a}'$ *and* $\mathbf{a}''$*, respectively.* For instance,

$$\begin{vmatrix} a + a' & b + b' \\ c & d \end{vmatrix} = (a + a')\,d - (b + b')\,c,$$

which can be rewritten as $ad - bc + a'd - b'c$, i.e.

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} + \begin{vmatrix} a' & b' \\ c & d \end{vmatrix}.$$

Of course, this property can be extended to the sum of any number of vectors. In particular:

(h) *If a line is the linear combination of other parallel lines, the determinant is* 0. This immediately follows from properties (f) and (g).

(i) *When adding a linear combination of the other lines to a given line of a matrix, the determinant is unchanged.* Again, this follows from properties (f) and (g).

(l) (**Binet's theorem**[17]) *The determinant of the matrix* $\mathbf{AB}$ *equals the product of the determinants:*

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

[17] Jacques Philippe Marie Binet (1786-1856), mechanic and astronomer.

The converse of property (h) can be proved to hold as well, i.e. if the determinant of a square matrix is zero, then the lines of the matrix are linearly dependent. We thus obtain:

Theorem 9.1. *The determinant of a square matrix is zero if and only if its rows and/or columns are linearly dependent.*

A technique for calculating determinants based on property (i), called the *reduction method*, is suitable for automatic computation. The idea is to build up, starting from a given matrix, another matrix with the same determinant, but with many more zero elements (e.g. a triangular matrix), so as to speed up the calculation of the determinant.

A last remark, quite useful for calculating the inverse matrix of a given matrix, is the following: *the sum of the products of the coefficients in a line multiplied by the algebraic complements of the corresponding elements of another parallel line is 0.* In formulae, for two rows: if $r \neq t$, then

$$\sum_{s=1}^{n} a_{rs} A_{ts} = 0$$

and analogously in the case of two columns.

8.10 Inverse matrix

In this section we shall answer the questions which arose at the end of Section 8. A necessary and sufficient condition for a square matrix $\mathbf{A}$ to be invertible is that its determinant be different from zero. In such a case there also exists a formula to compute the inverse matrix, which uses the transpose of the matrix of the algebraic complements. To be precise, if $\mathbf{A} = [a_{rs}]$, and A_{rs} is the algebraic complement of a_{rs}, the **adjoint** of $\mathbf{A}$ is the matrix

$$\mathbf{A}^* = [A_{sr}] = \begin{bmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ A_{12} & A_{22} & \dots & A_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1n} & A_{2n} & \dots & A_{nn} \end{bmatrix}$$

i.e. the transpose of the matrix obtained by replacing every element of $\mathbf{A}$ with its own algebraic complement. The following result holds.

Theorem 10.1. *A matrix $\mathbf{A}$ is invertible if and only if* $\det \mathbf{A} \neq 0$. *In such a case, its inverse matrix is given by the formula*

$$\mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} \mathbf{A}^*.$$

As a consequence,

$$\det \left(\mathbf{A}^{-1} \right) = \frac{1}{\det \mathbf{A}}.$$

In other words: a matrix is *singular* if and only if its determinant is zero.

We are not going to prove the theorem here. We limit ourselves to remarking that, if $\mathbf{A}^{-1}$ exists, then

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n.$$

so that, by calculating the determinant of both sides,

$$\det\left(\mathbf{A}^{-1}\mathbf{A}\right) = \det\left(\mathbf{A}\mathbf{A}^{-1}\right) = \det \mathbf{I}_n.$$

Recalling now Binet's Theorem, we obtain

$$\det\left(\mathbf{A}^{-1}\right) \cdot \det \mathbf{A} = \det \mathbf{A} \cdot \det\left(\mathbf{A}^{-1}\right) = 1. \tag{8.17}$$

Whence we deduce that $\det \mathbf{A}$ cannot be equal to 0 and that $\det\left(\mathbf{A}^{-1}\right) = 1/\det \mathbf{A}$.

Example 10.1. Let

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix}.$$

We obtain $\det \mathbf{A} = -5 \neq 0$, so $\mathbf{A}$ is invertible. The matrix of the algebraic complements of $\mathbf{A}$ is

$$\begin{bmatrix} 3 & -4 \\ -2 & 1 \end{bmatrix}$$

and its adjoint is

$$\mathbf{A}^* = \begin{bmatrix} 3 & -2 \\ -4 & 1 \end{bmatrix},$$

so that

$$\mathbf{A}^{-1} = -\frac{1}{5}\begin{bmatrix} 3 & -2 \\ -4 & 1 \end{bmatrix} = \begin{bmatrix} -3/5 & 2/5 \\ 4/5 & -1/5 \end{bmatrix}.$$

It is possible to check that the result is correct:

$$\begin{bmatrix} -3/5 & 2/5 \\ 4/5 & -1/5 \end{bmatrix} \cdot \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 4 & 3 \end{bmatrix} \cdot \begin{bmatrix} -3/5 & 2/5 \\ 4/5 & -1/5 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Example 10.2. We invite the reader to check the calculations, which are only partially shown below, in this second example with an order 3 matrix. Let

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \\ -1 & -1 & 0 \end{bmatrix}$$

Its determinant is -4, so $\mathbf{A}$ is invertible. By replacing the elements with their own algebraic complements, we obtain the matrix

$$\begin{bmatrix} 1 & -1 & -1 \\ -3 & 3 & -1 \\ -4 & 8 & -4 \end{bmatrix}.$$

For instance, the elements in the first row are obtained by means of the following calculations

$$A_{11} = \begin{vmatrix} 2 & 1 \\ -1 & 0 \end{vmatrix} = 1,\ A_{12} = -\begin{vmatrix} 3 & 1 \\ -1 & 0 \end{vmatrix} = -1,\ A_{13} = \begin{vmatrix} 3 & 2 \\ -1 & -1 \end{vmatrix} = -1.$$

The transpose of the above matrix is

$$\mathbf{A}^* = \begin{bmatrix} 1 & -3 & -4 \\ -1 & 3 & 8 \\ -1 & -1 & -4 \end{bmatrix}.$$

By dividing by $\det \mathbf{A}$, we obtain

$$\mathbf{A}^{-1} = -\frac{1}{4}\begin{bmatrix} 1 & -3 & -4 \\ -1 & 3 & 8 \\ -1 & -1 & -4 \end{bmatrix} = \begin{bmatrix} -1/4 & 3/4 & 1 \\ 1/4 & -3/4 & -2 \\ 1/4 & 1/4 & 1 \end{bmatrix}.$$

8.11 Rank of a matrix

We have seen that n vectors in $\mathbb{R}^n$ are linearly dependent if the (square) matrix obtained by matching them in rows or in columns is singular, i.e. if its determinant is zero. We now want to consider a problem which is slightly more general.

Consider n vectors $\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^n$ belonging to $\mathbb{R}^m$, where, in general, m is different from n. Let $C\left(\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^n\right)$ be the subspace of $\mathbb{R}^m$ generated by the vectors under consideration. We ask ourselves: *what is the dimension of* $C\left(\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^n\right)$?

To answer this question, we need to determine the maximum number of linearly *independent* vectors among $\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^n$. Since $\mathbb{R}^m$ has dimension m, such a number will surely be less than the minimum between m and n. To determine it precisely, we introduce the concept of the *rank of a matrix*.

Definition 11.1. *Let* $\mathbf{A}$ *be an* $m \times n$ *matrix. The* ***rank*** *or* ***characteristic*** *of* $\mathbf{A}$ *is the maximum number* $r \geq 0$ *of linearly independent columns in* $\mathbf{A}$.

It is indeed possible to show that the rank of a matrix $\mathbf{A}$ is the same as the maximum number of linearly independent *rows* in $\mathbf{A}$.

Example 11.1. Let

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 0 & 4 & 5 \\ -1 & -2 & 3 & 7 & 6 \\ 0 & 3 & 4 & 1 & -1 \end{bmatrix}.$$

It is straightforward to check that, for instance, the first three columns are linearly independent. Since there cannot be more than three linearly independent vectors in $\mathbb{R}^3$, we can then conclude that the rank of $\mathbf{A}$ is 3. It is also possible to verify that the three rows of $\mathbf{A}$ are indeed linearly independent.

The determination of the maximum number of linearly independent vectors in a set is quite a difficult task to implement for automatic computation. Nevertheless, determinants can be useful for determining the rank as well.

Suppose that the matrix $\mathbf{A}$ has rank r, and let $\mathbf{B}$ be the matrix obtained from r linearly independent columns of $\mathbf{A}$. Since the rank of $\mathbf{B}$ is still r, $\mathbf{B}$ has r linearly independent *rows* as well, or, in other words, it contains a square submatrix of order r which is not singular. Such a submatrix has, then, a non zero determinant – i.e. $\mathbf{B}$ (and hence $\mathbf{A}$, for which $\mathbf{B}$ is a submatrix) has a non-zero minor with order r. Conclusion:

Theorem 11.1. *Let $\mathbf{A}$ be an $m \times n$ matrix. The **rank** of $\mathbf{A}$ is the integer $r \geq 0$ such that: $\mathbf{A}$ has at least one non-zero minor of order r and all of the minors with order greater than r are zero.*

Example 11.2. Let us examine again the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 0 & 4 & 5 \\ -1 & -2 & 3 & 7 & 6 \\ 0 & 3 & 4 & 1 & -1 \end{bmatrix}.$$

Calculating the minor formed by the first three columns, we obtain

$$\begin{vmatrix} 1 & 2 & 0 \\ -1 & -2 & 3 \\ 0 & 3 & 4 \end{vmatrix} = 1 \cdot \begin{vmatrix} -2 & 3 \\ 3 & 4 \end{vmatrix} - 2 \cdot \begin{vmatrix} -1 & 3 \\ 0 & 4 \end{vmatrix} + 0 \cdot \begin{vmatrix} -1 & -2 \\ 0 & 3 \end{vmatrix} = -9,$$

and thus the rank of $\mathbf{A}$ is 3.

In this first example, we have been quite lucky. Indeed, had the determinant been equal to zero, how many other determinants should we calculate? In the above example, there are 10 minors of order 3! Had they been equal to zero, should we calculate them all?

Luckily, the answer is no. To determine the rank it is not necessary to take into consideration *all* of the minors of $\mathbf{A}$. Indeed, it is possible to prove the following result.

Theorem 11.2. *If the matrix $\mathbf{A}$ (of type $m \times n$) has a non-zero minor of order r ($r < \min\{m, n\}$) and if all the minors of order $r + 1$, which are obtained by "augmenting" such a minor with one row and one column of $\mathbf{A}$, are zero, then the rank of $\mathbf{A}$ is r.*

A calculation technique for determining the rank, called the *Kronecker's algorithm*, is based on the above theorem.

Example 11.3. Let

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 2 & 1 & -3 & 3 \\ 1 & -1 & 3 & 3 \end{bmatrix}.$$

Let us start by noticing that the rank of $\mathbf{A}$ is at least 2, since the minor

$$\begin{vmatrix} 1 & 0 \\ 2 & 1 \end{vmatrix}$$

is different from zero. Let us calculate the minors of order 3 which can be obtained by augmenting the non zero minor just found. Since both the minors

$$\begin{vmatrix} 1 & 0 & 0 \\ 2 & 1 & -3 \\ 1 & -1 & 3 \end{vmatrix} \quad \text{and} \quad \begin{vmatrix} 1 & 0 & 2 \\ 2 & 1 & 3 \\ 1 & -1 & 3 \end{vmatrix}$$

are zero, we conclude that the rank of $\mathbf{A}$ is 2.

The answer to the question asked at the beginning of the section is then the following: the dimension of the subspace $C\left(\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^n\right)$, spanned by the vectors $\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^n$, belonging to $\mathbb{R}^m$, is *the rank of the matrix*

$$\mathbf{A} = \left[\mathbf{a}^1\ \mathbf{a}^2 \cdots \mathbf{a}^n\right].$$

8.12 Exercises

8.1. A shop sells CDs, DVDs and stereo appliances. Calculate the weekly profit, knowing that the number of items sold, the unit sale prices and the unit costs for the shop are given by the following vectors:

$$\mathbf{q} = \begin{bmatrix} 700 \\ 400 \\ 20 \end{bmatrix}, \quad \mathbf{p} = \begin{bmatrix} 4 \\ 6 \\ 150 \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} 3 \\ 4 \\ 125 \end{bmatrix}.$$

8.2. Denoting by $\mathbf{1}$ the (row) vector of $\mathbb{R}^n$ with all of the components equal to 1, write the sum $\sum_{k=1}^{n} a_k$ as an inner product.

8.3. *True or false?*
(a) If a set of vectors contains the zero vector, they must be linearly dependent.
(b) Given a set of linearly dependent vectors, it is always possible to write any one of them as a linear combination of the other ones.
(c) Two non-zero vectors of $\mathbb{R}^3$ are always linearly independent.
(d) Four vectors of $\mathbb{R}^3$ are always linearly dependent.

8.4. Given the matrices

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & -1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 3 & 0 & 2 \\ -1 & 4 & 5 \end{bmatrix}$$

calculate, wherever possible, $\mathbf{A}^2$, $\mathbf{A}^{-1}$, $\mathbf{B}^{\mathrm{T}}$, $\mathbf{B}^{-1}$, $\mathbf{AB}$, $\mathbf{BA}$.

8.5. Calculate the volume of the parallelepiped built from the vectors

$$\mathbf{a} = (0, -1, 3), \quad \mathbf{b} = (1, 1, -2), \quad \mathbf{c} = (2, -3, 1).$$

Are the vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$ linearly independent?

8.6. Given the vectors

$$\begin{bmatrix} k \\ 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 4 \\ k \\ 1 \end{bmatrix}, \quad \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 3 \\ 1 \\ 0 \end{bmatrix},$$

determine how many of them are linearly independent, for every value of the parameter k in $\mathbb{R}$.

8.7. Many problems feature matrices:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{bmatrix},$$

such that the sum of the elements along each column is 1: $\sum_{r=1}^{n} p_{rs} = 1$, for every s (such matrices are called *stochastic* matrices). Check that $\mathbf{I}_n - \mathbf{P}$ is a singular matrix.

9

Linear Systems and Functions

Many applications in Economics involve dealing with problems where a large number of variables are linked together by rather simple (i.e. linear) relations. These problems naturally lead to the study of linear algebraic equations.

The appropriate language for this features vectors and matrices, which were covered in the last chapter. A system of linear equations (sometimes thousands of equations, involving thousands of unknown variables) may indeed be written as a single equation involving matrices and vectors. Such a concise, yet exhaustive, notation allows us to handle all of the expressions in a more convenient way. This is a remarkable point of strength, which helps to explain the success of these tools in economical analysis and in many management applications.

The chapter is organised as follows.

• Theories about the *existence*, *uniqueness* and *structure of the solutions* of a linear system are presented in the language of vectors and matrices. In particular, Cramer's theorem and Rouché-Capelli's theorem are covered.

• *Linear functions* from $\mathbb{R}^n$ to $\mathbb{R}^m$ are introduced. In such a setting, we reinterpret the results about the solvability of linear systems seen in the item above.

9.1 Linear systems

Let us start with a little terminology. An equation in the n unknowns $x_1, x_2, \ldots, x_n$ is called a *linear* equation if it can be written as

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

that is, if it is a polynomial equation of the first degree in the unknowns x_1, x_2, ..., x_n.

It is identified by the n real *coefficients* $a_1, a_2, \ldots, a_n$ and the *constant* b. Solving the equation means finding all of the n-tuples of real numbers $x_1, x_2, \ldots, x_n$ such that the equality is satisfied.

A *linear system* of m equations in the n unknowns $x_1, x_2, \ldots, x_n$ is built from m linear equations which are meant to be all satisfied at the same time:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \qquad\qquad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m. \end{cases} \tag{9.1}$$

The numbers a_{ij} $(1 \le i \le m, 1 \le j \le n)$ are called the *coefficients* of the system (a_{ij} is the coefficient of x_j in the i-th equation) and $b_1, b_2, \ldots b_m$ are the *constants*. When all of the b_is $(1 \le i \le m)$ are null, the system is called *homogeneous*.

A *solution* of the system is an n-tuple $x_1, x_2, \ldots, x_n$ of real numbers which satisfy all of the m equations simultaneously. A system admitting no solution is called *impossible*, otherwise it is called *possible* (or *solvable*). Two systems are called *equivalent* if they have the same solutions.

A question which is natural to ask when facing a linear system is: "*Is the system solvable*?". And, in the affirmative case, "*how many solutions are there, and how can they be determined*?".

If the system is simple, the solution can be found by means of elementary methods, as in the following example.

• (⇛ **Chapter 11**) *A small fiscal-financial problem.* Consider a short-term bond, paying 1000 Euro in 3 months.

Suppose at first its current price to be 975.61 Euro. We can then calculate its simple rate r of return by solving the equation

$$975.61\left(1 + r\frac{3}{12}\right) = 1000,$$

which yields $r = 0.1 = 10\%$.

Suppose now that, when purchasing a similar bond, a tax is paid, equal to 12.5% of the difference between the face value and the purchase price. Including such a tax, the title costs 990 Euro. We wonder what its gross return is, without taking the tax into account. To know the amount of tax to be deduced from the 990 Euro of the paid price, we would need to know the price P "before taxes", which is unknown. Let us call T the tax, which is unknown as well.

The values P and T are linked together by the relations

$$\begin{cases} P + T = 990 \\ T = 0.125\,(1000 - P)\,, \end{cases}$$

whence the price P and the tax T can be jointly determined. The system can be solved by means of the replacement method: T is deduced from the second equation

and replaced in the first, yielding the equivalent system

$$\begin{cases} P + 0.125\,(1000 - P) = 990 \\ T = 0.125\,(1000 - P)\,. \end{cases}$$

From the first equation, we get

$$P = \frac{990 - 125}{0.875} = 988.57.$$

Thus,

$$988.57\left(1 + r\frac{3}{12}\right) = 1000$$

and, finally, $r = 4.6249\%$.

9.1.1 Elimination method

When the number of equations and unknowns is large, we might prefer to use methods which are suitable for automatic computing, such as the *Gauss elimination method*, which we only briefly outline. Given a linear system, we want to obtain another linear system, equivalent to the starting one, by following (possibly many times) the following operations:

(a) shuffle the ordering of the equations;
(b) replace an equation with an equivalent one;
(c) add to some equation a linear combination of the other ones;
(d) remove from the system an equation which is a linear combination of the other ones.

A linear combination of equations of a system is an equation obtained by multiplying each of the involved equations by some non null constant, and then adding termwise. For instance, the equation

$$3(x + y) - 2(x - y) = 3 \cdot 5 - 2 \cdot 4,$$

which is equivalent to $x + 5y = 7$, is a linear combination of the equations $x + y = 5$ and $x - y = 4$ (with coefficients 3 and -2).

The *elimination method* allows for "transforming" a system into an equivalent one which is solvable (almost) at a glance. By exploiting the above operations which transform a system into another, equivalent one, it is indeed possible to reduce a system to a *triangular form*. This means manipulating a system in such a way that the first equation features all of the unknowns, one unknown (e.g., the first one) is missing from the second equation, the same unknown and one more are missing from the third equation, and so on. Let us provide an elementary example.

Example 1.1. Let us solve the system

$$\begin{cases} x + y + z = 1 \\ 2x + 3y - 4z = -3 \\ 3x - 4y + 5z = -2. \end{cases}$$

We can leave the first equation unchanged, and try and eliminate x from the other two equations. For instance, we replace the second equation with the difference between itself and the double of the first one (we may write II−2I) and the third equation with the difference between itself and the triple of the first (III−3I). We get

$$\begin{cases} x+y+z=1 & \text{I} \\ y-6z=-5 & \text{II}-2\text{I} \\ -7y+2z=-5 & \text{III}-3\text{I}. \end{cases}$$

Now, we leave the first two equations unchanged, and try and eliminate y from the third one. We can add to the third equation the second one multiplied by 7, thus getting

$$\begin{cases} x+y+z=1 & \text{I} \\ y-6z=-5 & \text{II} \\ -40z=-40 & \text{III}+7\text{II} \end{cases}$$

which is equivalent to

$$\begin{cases} x+y+z=1 \\ y-6z=-5 \\ z=1. \end{cases}$$

By replacing in the second equation the value we found for z, we can deduce the value for y and, consequently, get from the first equation the value for x. The system, then, has the unique solution

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

9.1.2 Linear systems and matrices

To deal in a general way with the study of a linear system such as (9.1), it is convenient to rewrite it as a single equation of the type

$$\mathbf{A}\mathbf{x}=\mathbf{b},$$

where $\mathbf{A}$ is the *matrix of the coefficients*, or simply the *coefficient matrix* (with m rows and n columns), $\mathbf{x}$ is the *vector of the unknowns* or the *unknown vector*, and $\mathbf{b}$ is the *vector of the constants* or the *constant vector*:

$$\mathbf{A}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix},\ \mathbf{x}=\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}\in\mathbb{R}^n,\ \mathbf{b}=\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}\in\mathbb{R}^m.$$

Example 1.12. For the system

$$\begin{cases} 3x_1+2x_2-x_3=1 \\ 12x_2-2x_3=10, \end{cases}$$

we have

$$\mathbf{A} = \begin{bmatrix} 3 & 2 & -1 \\ 0 & 12 & -2 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 10 \end{bmatrix}.$$

9.2 Systems with n equations and n unknowns

In this section we shall deal with systems featuring n equations and n unknowns, i.e., of the type

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{9.2}$$

where the coefficient matrix $\mathbf{A}$ is an order n square matrix.

In the case when $n = 1$, the matrix $\mathbf{A}$ and the vectors $\mathbf{x}$ and $\mathbf{b}$ reduce to real numbers, and the system reduces to the equation $ax = b$, a first degree equation in the only unknown x. If $a \neq 0$, the solution is unique and given by

$$x = b/a = a^{-1}b.$$

If $a = 0$, the equation is either impossible (if $b \neq 0$) or allows for infinitely many solutions (if $b = 0$, as it reduces to $0x = 0$).

Something similar is true for the general system (9.2) as well. Indeed, we have already seen that the condition $a \neq 0$, which allows us to write the inverse a^{-1} of a, is equivalent to the matrix condition $\det \mathbf{A} \neq 0$, which allows us to write the inverse matrix $\mathbf{A}^{-1}$ of $\mathbf{A}$. Analogously to the numeric case, when $\mathbf{A}$ is invertible the system (9.2) features existence and uniqueness of the solution, as the following *Cramer's*[1] *theorem* shows. Note that if $\det \mathbf{A} \neq 0$, then the matrices $\mathbf{A}$ and $[\mathbf{A}|\mathbf{b}]$ both have rank n, but we shall return to this remark later on.

Theorem 2.1. *Let $\mathbf{A}\mathbf{x} = \mathbf{b}$ be a system of n equations in n unknowns.*
If $\det \mathbf{A} \neq 0$, the system has a unique solution, given by the formula

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}. \tag{9.3}$$

Proof. If the determinant of $\mathbf{A}$ is not null, $\mathbf{A}$ is invertible. Thus, premultiplying both sides of (9.2) by $\mathbf{A}^{-1}$, we get

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

From the associative property, we can write $\left(\mathbf{A}^{-1}\mathbf{A}\right)\mathbf{x}$ instead of $\mathbf{A}^{-1}(\mathbf{A}\mathbf{x})$ and thus, since $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$ and $\mathbf{I}_n\mathbf{x} = \mathbf{x}$, we get $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.

The uniqueness of the solution follows from uniqueness of the inverse matrix. □

Cramer's rule

By using the algorithm we have seen in Section 10 of Chapter 8 to determine the inverse matrix, we can represent the solution (9.3) in a more explicit way. Indeed,

[1] Gabriel Cramer (1707-1752), a Swiss mathematician.

we obtain

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = \frac{1}{|\mathbf{A}|}\mathbf{A}^*\mathbf{b}.$$

It is now enough to recall the structure of $\mathbf{A}^*$. We know that the *rows* of the matrix $\mathbf{A}^*$ collect in an orderly way the algebraic complements of the elements of the *columns* of $\mathbf{A}$. Now, the component x_k of the solution vector is found through a product between the k-th row of $\mathbf{A}^*$ and the column $\mathbf{b}$. Because of Laplace's theorem, such a product coincides with the determinant of the matrix obtained by replacing the k-th column of $\mathbf{A}$ with the column $\mathbf{b}$ of the constants. The following formulae then hold true:

$$x_1 = \frac{\begin{vmatrix} b_1 & a_{12} & \dots & a_{1n} \\ b_2 & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_n & a_{n2} & \dots & a_{nn} \end{vmatrix}}{|\mathbf{A}|}, \quad x_2 = \frac{\begin{vmatrix} a_{11} & b_1 & \dots & a_{1n} \\ a_{21} & b_2 & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & b_n & \dots & a_{nn} \end{vmatrix}}{|\mathbf{A}|}, \quad \text{etc.},$$

which are sometimes written as

$$x_1 = \frac{|\mathbf{A}_1|}{|\mathbf{A}|}, \quad x_2 = \frac{|\mathbf{A}_2|}{|\mathbf{A}|}, \quad \text{etc.}$$

Example 2.1. Consider the simple system

$$\begin{cases} 3x_1 + 2x_2 = 5 \\ x_1 - x_2 = 0. \end{cases}$$

It has a unique solution, because the coefficient matrix is not singular:

$$\begin{vmatrix} 3 & 2 \\ 1 & -1 \end{vmatrix} = -3 - 2 = -5 \neq 0.$$

The solution is:

$$x_1 = \frac{\begin{vmatrix} 5 & 2 \\ 0 & -1 \end{vmatrix}}{-5} = \frac{-5}{-5} = 1, \qquad x_2 = \frac{\begin{vmatrix} 3 & 5 \\ 1 & 0 \end{vmatrix}}{-5} = \frac{-5}{-5} = 1.$$

9.3 General systems

Let us proceed to more general systems, with m equations in n unknowns.

To understand whether such a system is solvable, it is possible to prove an immediately operational condition. If $\mathbf{A} = \left[\mathbf{a}^1\ \mathbf{a}^2 \dots\ \mathbf{a}^n\right]$ is the coefficient matrix of the system, we call *complete matrix* the matrix (with m rows and $n+1$ columns) obtained by placing the constant column $\mathbf{b}$ alongside $\mathbf{A}$, i.e.,

$$[\mathbf{A}|\mathbf{b}] := \left[\mathbf{a}^1\ \mathbf{a}^2 \dots\ \mathbf{a}^n\ \mathbf{b}\right].$$

Let now r and r' be the ranks of $\mathbf{A}$ and $[\mathbf{A}|\mathbf{b}]$ respectively. The following *Rouché-Capelli's theorem*[2] holds:

[2] E. Rouché (1832-1910) - A. Capelli (1855-1910).

Theorem 3.1. *A necessary and sufficient condition for the system* $\mathbf{Ax} = \mathbf{b}$ *to be solvable is that* $r = r'$.

Proof. By emphasizing the columns of $\mathbf{A}$, instead of $\mathbf{Ax} = \mathbf{b}$ we can write

$$\mathbf{a}^1 x_1 + \mathbf{a}^2 x_2 + \cdots + \mathbf{a}^n x_n = \mathbf{b}, \tag{9.4}$$

which shows the vector $\mathbf{b}$ as a linear combination of the vectors $\mathbf{a}^1,\ \mathbf{a}^2, \ldots, \mathbf{a}^n$, taking the components of $\mathbf{x}$ as coefficients. The system, then, has solutions if and only if $\mathbf{b}$ is a linear combination of the columns of $\mathbf{A}$.

Now, the vector $\mathbf{b}$ is a linear combination of the columns of $\mathbf{A}$ if and only if the two sets $\left\{\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^n\right\}$ and $\left\{\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^n, \mathbf{b}\right\}$ of vectors contain the same number of linearly independent vectors (i.e., if and only if they span the same subspace of $\mathbb{R}^n$). This is equivalent to the equality between the two ranks r and r'. □

Example 3.1. Consider the following system, with three equations and three unknowns:

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 2 \\ 2 & 2 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ k \end{bmatrix}$$

where the coefficient matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 2 \\ 2 & 2 & 4 \end{bmatrix}$$

is singular. Let us study the solvability of such a system as k varies.

The coefficient matrix has rank $r = 2$. Indeed, its two uppermost rows are linearly independent (otherwise, one of them would be a multiple of the other), whereas the third one is twice the second. Let us now pass to the complete matrix:

$$[\mathbf{A}|\mathbf{b}] = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 1 & 1 & 2 & 2 \\ 2 & 2 & 4 & k \end{bmatrix}.$$

By exploiting Kronecker's algorithm, since the minor obtained with the first two rows and the first two columns is not null, we can limit ourselves to calculating

$$\begin{vmatrix} 1 & 0 & 1 \\ 1 & 1 & 2 \\ 2 & 2 & k \end{vmatrix} = k - 4 \tag{9.5}$$

which is null when $k = 4$. This means that

$$r' = \begin{cases} 2 & \text{if } k = 4 \\ 3 & \text{if } k \neq 4. \end{cases}$$

Therefore, if $k \neq 4$ the system is impossible.

9.3.1 Solution scheme

We now want to see how one can proceed to solve the system $\mathbf{Ax} = \mathbf{b}$ when $\mathbf{A}$ and $[\mathbf{A}|\mathbf{b}]$ have the same rank ($r = r'$). The method of solution is based upon the fact that the common value of the two ranks r and r' tells us *how many significant equations* there are in the system, that is how many linearly independent equations there are:

(a) Extract from the matrix $\mathbf{A}$ a *non null* order r minor.

(b) Take into consideration only the equations of the system which correspond to the r rows of the minor, and remove the other $m - r$ ones. As a matter of fact, the latter $m - r$ equations are automatically satisfied when the first r equations are.

(c) Keep the r unknowns, whose coefficients build up the columns of the selected minor, on the left hand side. The terms involving the other $n - r$ unknowns are to be taken to the right hand side (obviously changing the sign of all terms), and those unknowns will be considered as parameters.

(d) In this way, a system with r equations in r unknowns is obtained, which can be solved, e.g., by applying Cramer's rule or any other known method.

As a consequence of Cramer's theorem, if $n = r$ the system has a unique solution, whereas if $r < n$ we obtain solutions that depend on $n-r$ parameters (the unknowns which have been taken to the right hand side), which can assume arbitrary values. We thus obtain infinitely many solutions.

Let us consider example 3.1 again. When $k = 4$, we have

$$\begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 2 \\ 2 & 2 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}.$$

Since the minor built up from the first two rows and the first two columns is not null, the system is equivalent to

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

and has infinitely many solutions, which can be written in the form:

$$\begin{cases} x = -h + 1 \\ y = -h + 1 \\ z = h \end{cases} \qquad h \in \mathbb{R}.$$

Let us summarise what we have seen so far.

Let $\mathbf{Ax} = \mathbf{b}$ be a linear system with m equations in n unknowns, and let r and r', respectively, be the ranks of $\mathbf{A}$ and $[\mathbf{A}|\mathbf{b}]$:

if $r \neq r'$ the system is impossible;

if $r = r' = n$ the system has a unique solution;

if $r = r' < n$ the system has infinitely many solutions (depending on exactly $n-r$ parameters, sometimes called the "degrees of freedom" of the system).

9.4 Structure of the solutions

9.4.1 Homogeneous systems

A system where the constant vector is the null vector is called *homogeneous*. A homogeneous system is always possible, because indeed it always admits (at least) the null solution. For linear systems, the following theorem holds.

Theorem 4.1. *Let* $\mathbf{A}$ *be an* $m \times n$ *matrix with rank* r. *Then the set of solutions of the system* $\mathbf{Ax} = \mathbf{0}$ *is a subspace of* $\mathbb{R}^n$, *with dimension* $n - r$.

Proof. Call $\mathcal{N}$ the set of solutions of the system. We restrict ourselves to showing that $\mathcal{N}$ is a subspace of $\mathbb{R}^n$. The proof that its dimension is $n-r$ is left to the reader.

If $\mathcal{N} = \mathbf{0}$, this is evident[3]. In the other cases we need to show that, whenever $\mathbf{x}^1$ and $\mathbf{x}^2$ belong to $\mathcal{N}$ and α, β are real numbers, $\alpha\mathbf{x}^1 + \beta\mathbf{x}^2$ belongs to $\mathcal{N}$ as well.

Since $\mathbf{Ax}^1 = \mathbf{0}$ and $\mathbf{Ax}^2 = \mathbf{0}$, by multiplying the first equality by α and the second one by β, adding them together and applying the distributive property, we finally get $\mathbf{A}\left(\alpha\mathbf{x}^1 + \beta\mathbf{x}^2\right) = \mathbf{0}$, i.e., $\alpha\mathbf{x}^1 + \beta\mathbf{x}^2 \in \mathcal{N}$. □

In particular, if $r = n$ the system $\mathbf{Ax} = \mathbf{0}$ only admits the null solution, otherwise it has infinitely many solutions depending on $n - r$ parameters.

Example 4.1. Let us solve the system

$$\begin{bmatrix} 1 & 2 & 4 & -1 \\ 0 & -3 & -3 & 1 \\ 1 & -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Let us compute the rank of the coefficient matrix. Since the minor built up from the first two rows and the first two columns is different from 0, whereas

$$\begin{vmatrix} 1 & 2 & 4 \\ 0 & -3 & -3 \\ 1 & -1 & 1 \end{vmatrix} = 0 \qquad \text{and} \qquad \begin{vmatrix} 1 & 2 & -1 \\ 0 & -3 & 1 \\ 1 & -1 & 0 \end{vmatrix} = 0,$$

the rank r of the coefficient matrix of the system is 2. Only two equations in the system are "significant". Since the minor built up from the first two rows and the first two columns is different from 0, the system is equivalent to

$$\begin{cases} x_1 + 2x_2 = -4x_3 + x_4 \\ -3x_2 = 3x_3 - x_4. \end{cases}$$

By setting $x_4 = h$, $x_3 = k$, we get

$$\begin{cases} x_1 = h/3 - 2k \\ x_2 \ = h/3 - k \\ x_3 = k \\ x_4 = h \end{cases} \qquad h, k \in \mathbb{R}.$$

[3] Recall that the null vector alone gives a subspace of $\mathbb{R}^n$.

The solutions can be written in the vectorial form

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = h \begin{bmatrix} 1/3 \\ 1/3 \\ 0 \\ 1 \end{bmatrix} + k \begin{bmatrix} -2 \\ -1 \\ 1 \\ 0 \end{bmatrix} \qquad h, k \in \mathbb{R},$$

which highlights the fact that they build up a subspace $\mathcal{N}$ of $\mathbb{R}^4$ with dimension 2. It is the vector space spanned, for instance, by the vectors

$$\begin{bmatrix} 1/3 \\ 1/3 \\ 0 \\ 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -2 \\ -1 \\ 1 \\ 0 \end{bmatrix},$$

which constitute a basis of it, since they are linearly independent. All the other vectors in $\mathcal{N}$ are linear combinations of them.

9.4.2 Structure of the solutions of a linear system

Let us now deal with non-homogeneous systems.

Theorem 4.2. *Let* $\mathbf{x}^0$ *be a particular solution of the system* $\mathbf{Ax} = \mathbf{b}$. *Every other solution is given by*

$$\mathbf{x} = \mathbf{x}^0 + \mathbf{z} \tag{9.6}$$

where $\mathbf{z}$ *is a solution of the associated homogeneous system* $(\mathbf{Ax} = \mathbf{0})$.

Proof. Let us check that, if $\mathbf{x}^0$ is a solution of the system, i.e., if $\mathbf{Ax}^0 = \mathbf{b}$, every vector of the form (9.6) is a solution as well. Indeed, we get

$$\mathbf{Ax} = \mathbf{A}\left(\mathbf{x}^0 + \mathbf{z}\right) = \mathbf{Ax}^0 + \mathbf{Az} = \mathbf{b} + \mathbf{0} = \mathbf{b}.$$

Conversely, every solution $\mathbf{x}$ of the system can be written in the form (9.6). Indeed, from $\mathbf{Ax} = \mathbf{b}$ and $\mathbf{Ax}^0 = \mathbf{b}$, we get

$$\mathbf{A}\left(\mathbf{x} - \mathbf{x}^0\right) = \mathbf{Ax} - \mathbf{Ax}^0 = \mathbf{b} - \mathbf{b} = \mathbf{0}$$

and thus $\mathbf{x} - \mathbf{x}^0$ is a solution of the homogeneous system. By setting $\mathbf{x} - \mathbf{x}^0 = \mathbf{z}$, we get (9.6). □

Example 4.2. The system

$$\begin{bmatrix} 1 & 2 & 4 & -1 \\ 0 & -3 & -3 & 1 \\ 1 & -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

has the solution $\begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix}^{\mathrm{T}}$, as it is easy to check. Since we have already written the solutions of the corresponding homogeneous system (see example 4.1),

we can immediately conclude that the solutions of this system are:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} + h \begin{bmatrix} 1/3 \\ 1/3 \\ 0 \\ 1 \end{bmatrix} + k \begin{bmatrix} -2 \\ -1 \\ 1 \\ 0 \end{bmatrix} \qquad h, k \in \mathbb{R}.$$

9.5 Economic applications

• *Leontief model.* An economic system is structured in three sectors, which will be denoted by 1, 2 and 3 respectively. Think, for instance, of agriculture, industry and the tertiary sector. Every sector produces goods or services for the sector itself, for other sectors and for consumption. The agricultural sector, called *primary*, produces (among other things) corn, which is used in the same sector for seeding and in the industrial sector to produce brownies, afterwards meant for consumption. The industrial sector, called *secondary*, produces (among other things) vehicles, used in the same sector (cars, lorries, trailers...) or in the agricultural sector (tractors) or for services, and so on. Analogously, one could think of the relations between the services sector, called the *tertiary sector*, and the other two (agriculture and industry). For the sake of simplicity, we shall call "corn" the agricultural product, measured, for instance, in quintals, "vehicles" the industrial product and "advice", measured in "working days", the product of the tertiary sector.

To produce a unit of corn we need corn itself, vehicles and some advice (from a good agronomist). To produce a vehicle we need vehicles and advice. The following table shows the needs for each sector.

Sector → *Units needed* ↓	**Primary**	**Secondary**	**Tertiary**
corn	0.1	0	0.02
vehicles	0.01	0.1	0.1
advice	0.1	0.15	0

For instance, the number 0.02 in the top right corner means that every working day of an advisor requires 0.02 quintals of corn. Let us now take another table into consideration, gathering the amount of goods we want to be available for consumption

Remnant for consumption	**Primary**
corn	100
vehicles	10
advice	2

and let us ask how much every sector should produce to allow for such a remnant for the final use. To solve the problem we can set up a system of equations which specifies, for each sector, which part of the production is destined to final consumption and which to intermediate uses. Let us start by taking the agricultural sector,

and let x_1 be the amount of corn to be produced. Such a production volume needs to satisfy a "balance" equation of the type

$$x_1 = \text{ intermediate use} + \text{final use},$$

and the intermediate utilization of corn has to be in proportion with the volumes of production of the three sectors, according to the *technical coefficients* gathered in the first table seen above. The balance equation for x_1 is then

$$x_1 = 0.1x_1 + 0x_2 + 0.02x_3 + 100.$$

An analogous procedure yields the equation for the secondary sector

$$x_2 = 0.01x_1 + 0.1x_2 + 0.1x_3 + 10,$$

and, finally, for the tertiary sector as well

$$x_3 = 0.1x_1 + 0.15x_2 + 0x_3 + 2.$$

Gathering together the three equations in a system, we get

$$\begin{cases} x_1 = 0.1x_1 + 0x_2 + 0.02x_3 + 100 \\ x_2 = 0.01x_1 + 0.1x_2 + 0.1x_3 + 10 \\ x_3 = 0.1x_1 + 0.15x_2 + 0x_3 + 2 \end{cases}$$

which is equivalent to

$$\begin{cases} 0.9x_1 - 0.02x_3 = 100 \\ -0.01x_1 + 0.9x_2 - 0.1x_3 = 10 \\ -0.1x_1 - 0.15x_2 + x_3 = 2. \end{cases}$$

The solution is

$$\begin{cases} x_1 = 111.45 \\ x_2 = 14.044 \\ x_3 = 15.251. \end{cases}$$

The problem can be stated in a more general fashion. Suppose that n sectors produce some kind of goods. The vectors

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \text{and} \quad \mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}$$

gather, respectively, the productions of each sector and the final consumptions. The matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

contains the Leontief *technical coefficients* (a_{rs} represents the amount of goods of the r-th type needed to produce one unit of goods of the s-th type). The n balance equations are equivalent to the equation

$$\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{c}, \quad \text{i.e.,} \quad (\mathbf{I} - \mathbf{A})\,\mathbf{x} = \mathbf{c},$$

and typically the problem is to find $\mathbf{x}$, given $\mathbf{A}$ and $\mathbf{c}$. If the matrix $\mathbf{I} - \mathbf{A}$ is invertible, the balance equation

$$(\mathbf{I} - \mathbf{A})\,\mathbf{x} = \mathbf{c},$$

admits the unique solution

$$\mathbf{x} = (\mathbf{I} - \mathbf{A})^{-1}\,\mathbf{c}.$$

Such a solution is "acceptable" only if it is $\geq \mathbf{0}$, otherwise the economic system is called *non viable*. It is also possible to prove that, if $\mathbf{x} \geq \mathbf{0}$, then $\mathbf{x} \geq \mathbf{c}$ as well.

• *Markov chains with business applications.* Consider a market, where two firms (called 1 and 2) sell their products. Suppose that consumers are free to move, every week, from buying brand 1 to buying brand 2 or conversely. The *transition table* is the table gathering the percentages of customers moving from brand to brand and those that remain loyal.

Brands	**from 1**	**from 2**
towards 1	60%	30%
towards 2	40%	70%

This table tells us that, from week to week, 60% of brand 1 customers remain loyal, whereas 40% of them move to the other brand. Analogously, 70% of brand 2 customers remain with the same brand, and the remaining 30% move to brand 1. The "heart" of the table is the 2×2 matrix, called the *transition matrix*

$$\mathbf{T} = \begin{bmatrix} 0.6 & 0.3 \\ 0.4 & 0.7 \end{bmatrix}.$$

Naturally, every column must add up to one because the system is *closed*, i.e., no consumer enters or exits the market.

Suppose that in a certain week, which we shall call week 0, the customers are equally split (half and half) between the two brands, and that the total number of customers is 1000. The following table shows how they are distributed between the two brands.

Brand	**Customers**
1	500
2	500

.

From this table we can extract a "mathematical object", namely, the starting vector

$$\mathbf{x}^0 = \begin{bmatrix} 500 \\ 500 \end{bmatrix}.$$

The market repartition between the two firms in the following week can be easily calculated. Brand 1 will be bought by 60% of those already purchasing it plus 30% of those purchasing the other brand, now "fugitives" towards the first one:

$$60\% \cdot 500 + 30\% \cdot 500 = 450,$$

while brand 2 will be bought by the other 550 customers. Such a number can also be found as the sum of the "fugitives" of brand 1 (40% of 500) and customers faithful to brand 2 (70% of 500)

$$40\% \cdot 500 + 70\% \cdot 500 = 550.$$

It is interesting to note that it is possible to quickly obtain the new composition of the market, which is described by the vector

$$\mathbf{x}^1 = \begin{bmatrix} 450 \\ 550 \end{bmatrix},$$

as the product of the transition matrix $\mathbf{T}$ times the initial vector:

$$\mathbf{x}^1 = \mathbf{T}\mathbf{x}^0 = \begin{bmatrix} 0.6 & 0.3 \\ 0.4 & 0.7 \end{bmatrix} \begin{bmatrix} 500 \\ 500 \end{bmatrix} = \begin{bmatrix} 450 \\ 550 \end{bmatrix}.$$

To determine the distribution after two weeks, it is possible to compute the vector

$$\mathbf{x}^2 = \mathbf{T}\mathbf{x}^1 = \mathbf{T}\left(\mathbf{T}\mathbf{x}^0\right) = \mathbf{T}^2\mathbf{x}^0,$$

obtaining

$$\mathbf{x}^2 = \begin{bmatrix} 0.6 & 0.3 \\ 0.4 & 0.7 \end{bmatrix}^2 \begin{bmatrix} 500 \\ 500 \end{bmatrix} = \begin{bmatrix} 435 \\ 565 \end{bmatrix}.$$

In full generality, after n weeks, the composition of the market will be described by the vector

$$\mathbf{x}^n = \mathbf{T}^n\mathbf{x}^0.$$

We notice that, whatever the starting vector $\mathbf{x}^0$, if the two market segments communicate with each other the vector $\mathbf{x}^n$ can be seen to stabilize closer and closer towards a configuration

$$\mathbf{x}^* = \begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix},$$

which turns out to be independent of $\mathbf{x}^0$. Let us check this numerically. Consider the two starting points

$$\mathbf{x}^0 = \begin{bmatrix} 500 \\ 500 \end{bmatrix} \text{ already seen and } \mathbf{y}^0 = \begin{bmatrix} 200 \\ 800 \end{bmatrix}$$

and compute (with the help of a good *software*) the composition after 10 weeks

$$\mathbf{x}^{10} = \begin{bmatrix} 0.6 & 0.3 \\ 0.4 & 0.7 \end{bmatrix}^{10} \begin{bmatrix} 500 \\ 500 \end{bmatrix} \text{ and } \mathbf{y}^{10} = \begin{bmatrix} 0.6 & 0.3 \\ 0.4 & 0.7 \end{bmatrix}^{10} \begin{bmatrix} 200 \\ 800 \end{bmatrix}.$$

We get

$$\mathbf{x}^{10} = \begin{bmatrix} 429 \\ 571 \end{bmatrix} \quad \text{and} \quad \mathbf{y}^{10} = \begin{bmatrix} 429 \\ 571 \end{bmatrix}.$$

This property of independence with respect to the starting point is called *ergodicity*. Continuing the calculations, it is possible to see that

$$\mathbf{x}^{10} = \mathbf{y}^{10} = \mathbf{x}^{11} = \mathbf{y}^{11}.$$

The vector around which $\mathbf{x}^n$ stabilizes when n is big enough is called the *equilibrium vector*. How can an equilibrium vector $\mathbf{x}^*$ be determined? The idea is that, if the composition of the market were given by $\mathbf{x}^*$, after a week we should find it unchanged. In other words, we must have the condition

$$\mathbf{x}^* = \mathbf{T}\mathbf{x}^*, \tag{9.7}$$

which can be rewritten as

$$(\mathbf{I} - \mathbf{T})\,\mathbf{x}^* = \mathbf{0},$$

i.e.,

$$\left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0.6 & 0.3 \\ 0.4 & 0.7 \end{bmatrix} \right) \begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

This linear system is homogeneous, and thus solvable. It admits infinitely many solutions, because the determinant of the coefficients matrix

$$\det \begin{bmatrix} 0.4 & -0.3 \\ -0.4 & 0.3 \end{bmatrix}$$

is zero. This means that one of the two equations (say, the second one) is redundant, as it is dependent on the other. We can nevertheless replace it with the condition that the total number of customers be 1000,

$$x_1^* + x_2^* = 1000,$$

thus getting the new system

$$\begin{bmatrix} 0.4 & -0.3 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} = \begin{bmatrix} 0 \\ 1000 \end{bmatrix}$$

whose solution is

$$\begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix} = \begin{bmatrix} 429 \\ 571 \end{bmatrix},$$

the same we found above.

It is possible to take into consideration a market where there are $k \geq 2$ brands and N customers. In such a situation, analogous considerations apply: if it is always possible to shift from any chosen brand to another one (maybe not in a single transition, but possibly in several ones), then there exists a unique equilibrium configuration,

which is independent of the starting distribution and can be determined by adding to (9.7) the condition on the total size of the market

$$x_1^* + x_2^* + \cdots + x_k^* = N.$$

Then, as time goes by, the composition of the market gets indefinitely closer to such a configuration.

These models are known as *discrete Markov*[4] *chains*, and can be usefully applied in many other fields of study. They can be used, for instance, to model the evolution over time of a car fleet (see next example) or even to investigate the dynamics of voters' opinions (moving from a candidate to another during an election campaign). Markov chains are also used in Economic theory, for instance in studying income distribution. In particular, the ergodic property of these processes gives an explanation for the substantial stability over time and space of such distributions.

• *A car fleet at equilibrium.* A firm possesses several cars destined to the sales force. Any car can be in one of the following states: (1) in working order, (2) under repair and (3) under test. The following matrix $\mathbf{T}$ describes the percentages of cars moving from state to state which are observed on a weekly basis. The columns of $\mathbf{T}$ correspond to the origin state, while its rows correspond to the destination state.

$$\mathbf{T} = \begin{bmatrix} 90\% & 0 & 95\% \\ 9\% & 60\% & 5\% \\ 0 & 38\% & 0 \end{bmatrix}.$$

For instance, the 95% in the upper right corner means that 95% of the cars (3), i.e., under test, pass to state (1), i.e., in working order, after a week. If one wants to forecast the composition of the fleet in the next week based on the current state

$$\mathbf{x} = \begin{bmatrix} w \\ r \\ t \end{bmatrix}$$

with w working cars, r cars under repair and t cars under test, it is enough to perform the product $\mathbf{Tx}$. For instance, starting from

$$\mathbf{x} = \begin{bmatrix} 800 \\ 10 \\ 6 \end{bmatrix}$$

we get the forecast composition in the following week

$$\mathbf{Tx} = \begin{bmatrix} 90\% & 0 & 95\% \\ 9\% & 60\% & 5\% \\ 0 & 38\% & 0 \end{bmatrix} \begin{bmatrix} 800 \\ 10 \\ 6 \end{bmatrix} = \begin{bmatrix} 725.7 \\ 78.3 \\ 3.8 \end{bmatrix}.$$

[4] Such models were proposed, in a probabilistic setting, by the Russian mathematician Andrei A. Markov (1856-1922).

The situation after another week can be obtained with another multiplication:

$$\mathbf{T}^2\mathbf{x} = \begin{bmatrix} 656.74 \\ 112.48 \\ 29.754 \end{bmatrix}.$$

Note that the first two columns of $\mathbf{T}$ do not add up to one. This simply means that the percentage missing to reach 100% is the proportion of working cars, or cars under repair, which "exit" the system because of devastating crashes or because it is not worth repairing them. In such a situation, it is natural to think about a plan of regular substitution, consisting for instance in buying 10 new cars every week (of course in working order). This means that the population in a given week is no longer obtained by what is left of the existing fleet; to such a composition the new cars have to be added:

$$\mathbf{b} = \begin{bmatrix} 10 \\ 0 \\ 0 \end{bmatrix}.$$

It is reasonable to wonder whether the "car fleet" system features a time invariant equilibrium $\mathbf{x}^*$. It should be

$$\mathbf{x}^* = \mathbf{T}\mathbf{x}^* + \mathbf{b}, \quad \text{i.e.,} \quad (\mathbf{I} - \mathbf{T})\mathbf{x}^* = \mathbf{b}$$

whence, if $\mathbf{I} - \mathbf{T}$ is not singular,

$$\mathbf{x}^* = (\mathbf{I} - \mathbf{T})^{-1}\,\mathbf{b}.$$

In our case, we get

$$\mathbf{I} - \mathbf{T} = \begin{bmatrix} 10\% & 0 & -95\% \\ -9\% & 40\% & -5\% \\ 0 & -62\% & 0 \end{bmatrix}$$

whence

$$(\mathbf{I} - \mathbf{T})^{-1} = \begin{bmatrix} 67.914 & 64.349 & 67.736 \\ 16.043 & 17.825 & 16.132 \\ 6.0963 & 6.7736 & 7.1301 \end{bmatrix},$$

and so the equilibrium situation is

$$\mathbf{x}^* = (\mathbf{I} - \mathbf{T})^{-1}\,\mathbf{b} = \begin{bmatrix} 679.14 \\ 160.43 \\ 60.963 \end{bmatrix}.$$

Again, it is possible to prove that the equilibrium configuration attracts the state of the system whatever the starting position: indeed, such a state keeps on getting indefinitely closer to the equilibrium position. This fact has many implications in terms of resource management. Given the matrix $\mathbf{T}$, a natural consequence of the internal features of the system is that it is possible to control the system with an exogenous choice of the inputs.

If we call b the number of new cars bought weekly, the equilibrium state of the system is described by the vector

$$(\mathbf{I} - \mathbf{T})^{-1} b \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

If, for instance, we want the expected number of working cars to be f^*, it is sufficient to evaluate the first component of such a vector, i.e.,

$$67.914 \cdot b,$$

and require it to be equal to f^*. We then find $b = f^*/67.914 = f^* \cdot 0.014725$. In brief, if we want 800 working cars at equilibrium, we should buy every week $800 \times 0.014725 = 11.78$ new cars.

9.6 Linear functions from $\mathbb{R}^n$ to $\mathbb{R}^m$

If we multiply an $m \times n$ type matrix $\mathbf{A}$ by a vector $\mathbf{x}$ of $\mathbb{R}^n$, we get a vector $\mathbf{y}$ of $\mathbb{R}^m$. Then, to every vector $\mathbf{x}$ of $\mathbb{R}^n$ we can associate the vector $\mathbf{y}$ of $\mathbb{R}^m$ obtained this way:

$$\mathbf{x} \longmapsto \mathbf{y} = \mathbf{A}\mathbf{x},$$

thus obtaining a function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$. Let us see an example.

• *Use of machines in a firm.* A production process requires the use of two machines (a smoother and a cleaner). There are three types of products which can be obtained (luxury, average and *economy*), and each of them requires different operation times for the two machines. The following table shows the necessary times for each product and each machine.

minutes/item ↓; product →	**luxury**	**average**	**economy**
smoother	15	10	5
cleaner	6	4	3

Denote by x_1, x_2, x_3 the number of items produced for each type (x_1 the luxury, x_2 the average, x_3 the *economy* ones). The overall machine time necessary for smoothing is

$$y_1 = 15x_1 + 10x_2 + 5x_3.$$

It is a number y_1, function of the three variables x_1, x_2, x_3. If the number of items produced were, respectively, 100, 120 and 140, the machine time would be (in minutes)

$$y_1 = 15 \cdot 100 + 10 \cdot 120 + 5 \cdot 140 = 3400.$$

The analogous need for cleaning defines another function of three variables

$$y_2 = 6x_1 + 4x_2 + 3x_3,$$

which, corresponding to the production vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 100 \\ 120 \\ 140 \end{bmatrix},$$

defines a cleaning time (in minutes) of

$$y_2 = 6 \cdot 100 + 4 \cdot 120 + 3 \cdot 140 = 1500.$$

We can thus represent the dependence of the vector $\mathbf{y}$, gathering the operation times and having dimension $m = 2$, with respect to the production vector $\mathbf{x}$, having dimension $n = 3$, by means of the formula

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 15x_1 + 10x_2 + 5x_3 \\ 6x_1 + 4x_2 + 3x_3 \end{bmatrix}, \tag{9.8}$$

which can be written as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 15 & 10 & 5 \\ 6 & 4 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix},$$

or, more concisely, $\mathbf{y} = \mathbf{A}\mathbf{x}$, where we have set

$$\mathbf{A} = \begin{bmatrix} 15 & 10 & 5 \\ 6 & 4 & 3 \end{bmatrix}.$$

The system of *input-output* (i.e., production-machine time) relations, then, amounts to the multiplication by a matrix. Note that the function which associates the vector $\mathbf{y}$ (machine times) to the vector $\mathbf{x}$ (production) is *additive*, i.e., given two production vectors $\mathbf{x}$ and $\mathbf{x}'$ with respective required times $\mathbf{f}(\mathbf{x})$ and $\mathbf{f}(\mathbf{x}')$, the time $\mathbf{f}(\mathbf{x} + \mathbf{x}')$ required to obtain $\mathbf{x} + \mathbf{x}'$ is the sum of the two required times we started with. Analogously, $\mathbf{f}$ is *homogeneous*, i.e., when multiplying all of the production quantities by the same amount α (e.g., $\alpha = 2$ to double production, $\alpha = 1/2$ to halve it, ...) the required machine times get multiplied by the same factor.

Of course it is possible to generalise the framework to the case with n products and m times, so that $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$ and $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$. The law $\mathbf{f}$, associating to the generic production basket the corresponding vector of required machine times, will obey an expression such as:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$

where, for instance, a_{23} represents the unit item time for the second machine and the third product. In a compact form, it can still be written as $\mathbf{y} = \mathbf{A}\mathbf{x}$, where

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}$$

is the matrix gathering the required unit times (still called technical coefficients).

The function $\mathbf{f}$ satisfies the property of *linearity*, i.e., for every pair of vectors $\mathbf{x}$ and $\mathbf{x}'$ and for every real number α, we have

$$\begin{aligned} \mathbf{f}(\mathbf{x}+\mathbf{x}') &= \mathbf{f}(\mathbf{x}) + \mathbf{f}(\mathbf{x}') \qquad \textit{additivity} \\ \mathbf{f}(\alpha\mathbf{x}) &= \alpha\mathbf{f}(\mathbf{x}) \qquad \textit{homogeneity (of degree 1)}. \end{aligned}$$

Definition 6.1. *An additive and homogeneous function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called a* ***linear*** *function.*

These two properties can be put together: a function $\mathbf{f}$ is linear if and only if for every pair of real numbers α, β and for every pair of vectors $\mathbf{x}, \mathbf{x}'$:

$$\mathbf{f}(\alpha\mathbf{x}+\beta\mathbf{x}') = \alpha\mathbf{f}(\mathbf{x}) + \beta\mathbf{f}(\mathbf{x}').$$

Instead of linear function, the term *linear map* can be used. According to our previous considerations, every formula of the type

$$\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}, \tag{9.9}$$

with $\mathbf{A}$ an $m \times n$ type matrix, defines a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$.

Actually, all linear functions belong to this type, meaning that it is possible to find a matrix which "represents" them. Indeed:

Theorem 6.1. *A function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ is linear if and only if there exists a matrix $\mathbf{A}$ such that*

$$\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}.$$

Such a matrix is unique once the (canonical) bases have been fixed in $\mathbb{R}^n$ and $\mathbb{R}^m$, and is called the ***representing*** *matrix of the function $\mathbf{f}$.*

Proof. If $\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$, the known properties of matrix products imply that, for every $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^n$ and every $\alpha, \beta \in \mathbb{R}$, we have

$$\mathbf{f}(\alpha\mathbf{x}+\beta\mathbf{x}') = \mathbf{A}(\alpha\mathbf{x}+\beta\mathbf{x}') = \alpha\mathbf{A}\mathbf{x}+\beta\mathbf{A}\mathbf{x}' = \alpha\mathbf{f}(\mathbf{x}) + \beta\mathbf{f}(\mathbf{x}'),$$

and thus $\mathbf{f}$ is linear.

Conversely, let $\mathbf{f}$ be linear. Take the fundamental vectors of $\mathbb{R}^n$

$$\mathbf{e}^1,\ \mathbf{e}^2, \dots,\ \mathbf{e}^n$$

and call $\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^n$ their images with respect to $\mathbf{f}$, i.e., $\mathbf{a}^s = \mathbf{f}(\mathbf{e}^s)$ for every $s = 1, \ldots, n$. We know that, if $\mathbf{x} \in \mathbb{R}^n$, it is possible to write

$$\mathbf{x} = \sum_{s=1}^{n} x_s \mathbf{e}^s,$$

and thus

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}\left(\sum_{s=1}^{n} x_s \mathbf{e}^s\right) = \sum_{s=1}^{n} x_s \mathbf{f}(\mathbf{e}^s) = \sum_{s=1}^{n} \mathbf{a}^s x_s = \left[\mathbf{a}^1\ \mathbf{a}^2 \ldots\ \mathbf{a}^n\right] \mathbf{x} = \mathbf{A}\mathbf{x}.$$

Summarising, the matrix $\mathbf{A}$, which is obtained by setting the columns equal to the image vectors of the fundamental vectors of $\mathbb{R}^n$, is associated to the linear map $\mathbf{f}$. □

Example 6.1. The generic linear maps from $\mathbb{R}^2$ to $\mathbb{R}$ and from $\mathbb{R}^3$ to $\mathbb{R}$ are:

$$(x_1, x_2) \mapsto a_1 x_1 + a_2 x_2,$$
$$(x_1, x_2, x_3) \mapsto a_1 x_1 + a_2 x_2 + a_3 x_3.$$

Example 6.2. Let $\mathbf{f} : \mathbb{R}^3 \to \mathbb{R}^2$ be defined by the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & -1 \\ 0 & 2 & 4 \end{bmatrix}.$$

Multiplying $\mathbf{A}$ by $\mathbf{e}^1$, yields

$$\mathbf{A}\mathbf{e}^1 = \begin{bmatrix} 1 & 1 & -1 \\ 0 & 2 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and, analogously,

$$\mathbf{A}\mathbf{e}^2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad \mathbf{A}\mathbf{e}^3 = \begin{bmatrix} -1 \\ 4 \end{bmatrix}.$$

We see clearly, therefore, that the function maps the three vectors of the canonical basis of $\mathbb{R}^3$ precisely into the three columns of $\mathbf{A}$.

Composite function and matrix product

Given two linear functions $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{g} : \mathbb{R}^m \to \mathbb{R}^p$, respectively represented by the matrices $\mathbf{A}$ (of type $m \times n$) and $\mathbf{B}$ (of type $p \times m$), it is possible to define the *composite function* $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^p$, denoted by $\mathbf{g} \circ \mathbf{f}$ and defined by consecutively applying $\mathbf{f}$ and $\mathbf{g}$:

$$\mathbf{x} \mapsto \mathbf{y} = \mathbf{f}(\mathbf{x}) \mapsto \mathbf{z} = \mathbf{g}(\mathbf{y}) = \mathbf{g}[\mathbf{f}(\mathbf{x})] = (\mathbf{g} \circ \mathbf{f})\mathbf{x}.$$

The $(p \times n)$ matrix associated with the composite function $\mathbf{h} = \mathbf{g} \circ \mathbf{f}$ is precisely the product $\mathbf{BA}$. Indeed,

$$\mathbf{x} \mapsto \mathbf{y} = \mathbf{A}\mathbf{x} \mapsto \mathbf{z} = \mathbf{B}\mathbf{y} = \mathbf{B}(\mathbf{A}\mathbf{x}) = (\mathbf{B}\mathbf{A})\mathbf{x}.$$

• *A two-step production process.* A firm uses the physical amounts

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^{\mathrm{T}}$$

of n raw materials to obtain the physical amounts

$$\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_m \end{bmatrix}^{\mathrm{T}}$$

of m semifinished products, according to the linear function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$

$$\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x},$$

describing the first step of the production process. Such semifinished products feed the production of p final goods in the amounts

$$\mathbf{z} = \begin{bmatrix} z_1 & z_2 & \cdots & z_p \end{bmatrix}^{\mathrm{T}},$$

described by the linear function $\mathbf{g} : \mathbb{R}^m \to \mathbb{R}^p$

$$\mathbf{g}(\mathbf{y}) = \mathbf{B}\mathbf{y},$$

which summarises the second step of the process. The dependence of the final products on the raw materials is described by the linear function $\mathbf{g} \circ \mathbf{f} : \mathbb{R}^n \to \mathbb{R}^p$,

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

Inverse function and inverse matrix

Given a linear function $\mathbf{f}: \mathbb{R}^n \to \mathbb{R}^m$, we can easily prove that $\mathbf{f}$ is invertible if and only if the matrix $\mathbf{A}$ representing $\mathbf{f}$ is square (this means that it has to be $m = n$) and invertible, which happens if and only if $\det \mathbf{A} \neq 0$.

Moreover, if a linear function $\mathbf{f}$ is invertible and its representing matrix is $\mathbf{A}$, then its inverse function $\mathbf{f}^{-1}$ is also linear and it is represented by the matrix $\mathbf{A}^{-1}$. Indeed, from $\mathbf{y} = \mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$, if $\det \mathbf{A} \neq 0$, a left multiplication by $\mathbf{A}^{-1}$ shows that

$$\mathbf{x} = \mathbf{f}^{-1}(\mathbf{y}) = \mathbf{A}^{-1}\mathbf{y}.$$

Affine linear functions

We call *affine linear* the functions which are obtained by "translating" a linear function. A affine linear function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ is defined by a formula of the type

$$\mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b},$$

where $\mathbf{A}$ is an $m \times n$ type matrix, and $\mathbf{b}$ is a vector of $\mathbb{R}^m$.

Example 6.2. The function $\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{f}(x_1, x_2) = \begin{bmatrix} -2 & 1 \\ -3 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ -2 \end{bmatrix}$$

is obtained by translating the linear function

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{g}(x_1, x_2) = \begin{bmatrix} -2 & 1 \\ -3 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

by means of the vector $\begin{bmatrix} 1 \\ -2 \end{bmatrix}$.

9.6.1 Image and kernel of a linear function

The fact that *the linear functions* $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ *coincide with the functions* $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ *which can be described by* (9.9) allows us to express in a new language all of the concepts and theorems seen in the previous sections. In what follows, it is convenient to think that the bases in $\mathbb{R}^n$ and $\mathbb{R}^m$ are fixed, and to identify a linear function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ with its representing matrix $\mathbf{A}$, of type $m \times n$.

Definition 6.2. *The* ***image set*** *(or simply* ***image****) of* $\mathbf{A}$ *is the subset of* $\mathbb{R}^m$

$$\operatorname{Im} \mathbf{A} = \{\mathbf{y} \in \mathbb{R}^m : \ \mathbf{y} = \mathbf{A}\mathbf{x}, \ \mathbf{x} \in \mathbb{R}^n\}.$$

For every $\mathbf{x} \in \mathbb{R}^n$, the vector $\mathbf{A}\mathbf{x} \in \mathbb{R}^m$ is a linear combination of the columns $\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^n$, i.e.,

$$\mathbf{A}\mathbf{x} = \mathbf{a}^1 x_1 + \mathbf{a}^2 x_2 + \ldots + \mathbf{a}^n x_n.$$

As $\mathbf{x}$ varies, we obtain *all* of the linear combinations of the columns of $\mathbf{A}$, i.e., the linear space (indeed, a subspace of $\mathbb{R}^m$) spanned by the columns of $\mathbf{A}$.

Theorem 6.2. *The image* $\operatorname{Im} \mathbf{A}$ *is the subspace of* $\mathbb{R}^m$ *spanned by the columns of* $\mathbf{A}$*. Its dimension is equal to the rank of* $\mathbf{A}$.

To understand whether a vector $\mathbf{b}$ taken in $\mathbb{R}^m$ belongs to the image of $\mathbf{A}$ or not, we thus need to check whether it can be expressed as a linear combination of the columns of $\mathbf{A}$. According to what we have seen in the previous chapter, this amounts to checking whether the two matrices $\mathbf{A}$ and $[\mathbf{A}|\mathbf{b}]$ have the same rank. This allows us to "restyle" Rouché-Capelli's theorem in the language of linear maps.

Theorem 6.3. *A vector* $\mathbf{b} \in \mathbf{R}^m$ *belongs to* $\operatorname{Im} \mathbf{A}$ *if and only if the ranks of the two matrices* $\mathbf{A}$ *and* $[\mathbf{A}|\mathbf{b}]$ *coincide.*

When dealing with a linear map $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$, it is often interesting to determine which vectors in $\mathbb{R}^n$ get mapped into the null vector in $\mathbb{R}^m$. The set of such vectors is called the *kernel* of $\mathbf{A}$ and is usually denoted by $\ker \mathbf{A}$.

Definition 6.3. *The* ***kernel*** *of* $\mathbf{A}$ *is the subset of* $\mathbb{R}^n$

$$\ker \mathbf{A} := \{\mathbf{x} \in \mathbb{R}^n : \ \mathbf{A}\mathbf{x} = \mathbf{0}\}.$$

• (⇛ **Chapter 11**) *Net Present Value.* The additivity and homogeneity properties, which are satisfied by linear maps, have a precise economical and financial interpretation. Note that the *Net Present Value* (*NPV*) of a financial operation described by the cash flow gathered in the vector

$$\mathbf{x} = \begin{bmatrix} x_0 & x_1 & \cdots & x_n \end{bmatrix}^{\mathrm{T}},$$

with respect to the discount factors gathered in the vector

$$\boldsymbol{\psi} = \begin{bmatrix} \psi_0 & \psi_1 & \cdots & \psi_n \end{bmatrix},$$

turns out to be

$$\boldsymbol{\psi}\mathbf{x}.$$

Given the discount factors, then, the net present value is a linear function of the cash flow to be evaluated. If we enter into the two financial operations $\mathbf{x}$ and $\mathbf{x}'$, yielding an overall cash flow $\mathbf{x} + \mathbf{x}'$, the net present value of the portfolio encompassing both operations is the sum of the two net present values. Analogously, if instead of the operation described by $\mathbf{x}$ we enter into the multiple operation $c\mathbf{x}$, with c a real number, e.g. $c = 2$, the net present value of the operation gets multiplied by the same constant c, i.e. it doubles if $c = 2$.

Let now $\mathbf{x}$ be an investment with net present value $\boldsymbol{\psi}\mathbf{x}$. We wonder which financial operations $\mathbf{y}$ (for instance, loans to support the investment) are such that they neither decrease nor increase the Net Present Value of the investment. We need that

$$\boldsymbol{\psi}(\mathbf{x} + \mathbf{y}) = \boldsymbol{\psi}\mathbf{x},$$

whence, by the distributive property, $\boldsymbol{\psi}\mathbf{x} + \boldsymbol{\psi}\mathbf{y} = \boldsymbol{\psi}\mathbf{x}$, and then

$$\boldsymbol{\psi}\mathbf{y} = 0.$$

Thus, the financial operations $\mathbf{y}$ we are looking for are those which get mapped into 0 by $\boldsymbol{\psi}$. In financial language, they are said to have null Net Present Value.

The kernel of $\mathbf{A}$ is composed of the solutions of the homogeneous system $\mathbf{Ax} = \mathbf{0}$ and therefore, as we have shown in Section 4.1:

Theorem 6.4. *The kernel of a matrix* $\mathbf{A}$ *of type* $m \times n$ *is a subspace of* $\mathbb{R}^n$ *with dimension* $n - r$, *where* r *is the rank of* $\mathbf{A}$.

Let us allow ourselves another *restyling*. We have seen that the solutions of a linear system $\mathbf{Ax} = \mathbf{b}$ are obtained by adding to a particular solution of it all of the solutions of the associated homogeneous system. In terms of linear functions, the statement becomes:

Theorem 6.5. *All of the inverse images of a given vector* $\mathbf{b} \in \operatorname{Im}\mathbf{A}$ *are obtained from one of them by adding all of the elements in* $\ker\mathbf{A}$.

Example 6.3. Let

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$

and thus

$$\mathbf{A}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 + x_2 \\ x_1 + x_2 \end{bmatrix}.$$

The kernel of $\mathbf{A}$ is the set of all those vectors in $\mathbb{R}^2$ such that $x_1 + x_2 = 0$, which geometrically corresponds to the points of the straight line (passing through the

origin) whose equation is $x_2 = -x_1$. It is a subspace of $\mathbb{R}^2$, as such vectors can be written in the form

$$\mathbf{z} = h \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad h \in \mathbb{R}.$$

Consider now the vector $\mathbf{x}^0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, whose image is

$$\mathbf{y} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}.$$

The vectors $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ which have the same image as $\mathbf{x}^0 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ are such that

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix},$$

whence

$$\begin{cases} x_1 + x_2 = 3 \\ x_1 + x_2 = 3. \end{cases}$$

Setting $x_1 = 1 + h$ with $h \in \mathbb{R}$, we get

$$\begin{cases} x_1 = 1 + h \\ x_2 = 2 - h. \end{cases}$$

This shows that all inverse images of the vector $\mathbf{y} = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$ are the vectors

$$\mathbf{x} = \mathbf{x}^0 + \mathbf{z} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 1 \\ -1 \end{bmatrix} h, \qquad h \in \mathbb{R},$$

in agreement with theorem 6.5.

Recalling theorems 6.2 and 6.4, we can conclude that the dimensions of the *kernel* and of the *image* of a given linear map $\mathbf{f}\colon \mathbb{R}^m \to \mathbb{R}^n$ are connected by the following formula:

$$\boxed{\dim \operatorname{Im} \mathbf{A} + \dim \ker \mathbf{A} = n.}$$

9.7 Exercises

9.1. *True or false?*

(*a*) If $\mathbf{A}$ is a 7×6 type matrix, with maximum rank, then the system $\mathbf{Ax} = \mathbf{b}$ is always impossible.

(*b*) If $\mathbf{A}$ is a 6×7 type matrix, with maximum rank, then the system $\mathbf{Ax} = \mathbf{b}$ is never impossible.

9.2. Determine, as the parameter a varies, the number of solutions of the system

$$\begin{bmatrix} 1 & a & 2 \\ a & 9 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ -a \\ 0 \end{bmatrix}.$$

9.3. Solve the system

$$\begin{cases} 2x + y - z + 4t = 5 \\ x + y + z = 1 \\ x + 2z - t = 1. \end{cases}$$

9.4. Solve the system, as $k, h \in \mathbb{R}$ vary:

$$\begin{bmatrix} 1 & 0 & 2 \\ 1 & 1 & 2 \\ 2 & 1 & 4 \\ 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ k \\ h \end{bmatrix}$$

9.5. In a viable Leontief model (its solutions $\mathbf{x}$ are always ≥ 0) the level of final consumptions is null for every type of goods. Is it true that the the system must not produce anything? Or is it necessary for it to produce something, in order to remain viable?

9.6. In an economic system, the value of two economic policy targets y_1, y_2 can be reached simultaneously by appropriately choosing three instruments x_1, x_2, x_3, according to the relations scheme:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

Determine the choices of instruments which consent both targets to be kept as null.

9.7. In a Markov model, the brand-to-brand transition matrix is

$$\mathbf{P} = \begin{bmatrix} 0.5 & 0 & 0.2 \\ 0.25 & 1 & 0.1 \\ 0.25 & 0 & 0.7 \end{bmatrix}.$$

Columns correspond to starting brands, and rows correspond to target brands. Calculate $\mathbf{P}^2$ and provide an interpretation for it. Write the relations which must hold among the three market quotas $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^{\mathrm{T}}$ for the system to reach a stable distribution. Deduce $\mathbf{x}$ supposing that there are 1000 customers. Explain the result intuitively.

9.8. A community is made up of a number y of young people who are not employed, a number w of workers and a number e of retired people. Yearly transition percentages from young people to workers and from workers to retired people are, respectively, 5% and 10%. Among the retired population, the annual death rate is 15%. Suppose the annual death rate among young people and workers is null. The vector

$$\mathbf{x}(t) := \begin{bmatrix} y(t) \\ w(t) \\ e(t) \end{bmatrix}$$

describes the population at the beginning of a certain year t. Build the transition matrix $\mathbf{T}$ which, applied to such a vector, gives the composition at the beginning of the subsequent year $t + 1$.

Suppose now that γ young people enter the community every year. Write the recurrence relation between vectors $\mathbf{x}(t+1)$ and $\mathbf{x}(t)$, including the entries vector $\mathbf{b} = \begin{bmatrix} \gamma & 0 & 0 \end{bmatrix}^{\mathrm{T}}$. Show that such a population has an equilibrium configuration $\mathbf{x}^* = \begin{bmatrix} y^* & w^* & e^* \end{bmatrix}^{\mathrm{T}}$ (i.e., $\mathbf{x}(t+1) = \mathbf{x}(t)$) and compute it. Deduce the ratio between the equilibrium number of workers and retired people.

9.9. Determine the dimension of the image and of the kernel of the linear function represented by the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 0 & -1 \\ 0 & 4 & 3 & -2 \end{bmatrix}.$$

9.10. Let $\mathbf{f} : \mathbb{R}^4 \to \mathbb{R}^3$ be a linear function such that

$$\begin{aligned} f(\mathbf{e}^1) &= \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}, \qquad f(\mathbf{e}^2) = \begin{pmatrix} 0 \\ 2 \\ -1 \end{pmatrix}, \\ f(\mathbf{e}^3) &= f(\mathbf{e}^1) + 2f(\mathbf{e}^2), \quad f(\mathbf{e}^4) = -f(\mathbf{e}^3) \end{aligned}$$

($\mathbf{e}^1$, $\mathbf{e}^2$, $\mathbf{e}^3$, $\mathbf{e}^4$ denote the fundamental vectors in $\mathbb{R}^4$).

(a) Write the matrix $\mathbf{A}$ associated with $\mathbf{f}$.

(b) Determine a basis for $\ker \mathbf{f}$.

9.11. In some economic policy models, one assumes that m government targets $\mathbf{y} \in \mathbb{R}^m$ have to be achieved by choosing n instruments $\mathbf{x} \in \mathbb{R}^n$, linked together by the relation $\mathbf{y} = \mathbf{M}\mathbf{x}$, with $\mathbf{M}$ an $m \times n$ matrix of given coefficients. Assume $m \neq n$. Is it true that:

(a) if all of the instruments are chosen to be null, there are some special matrices $\mathbf{M}$ such that some target is not null?

(b) if all of the components of $\mathbf{x}$ are changed by a percentage k, then all of the components of $\mathbf{y}$ get changed by the same percentage?

(c) for some special matrices $\mathbf{M}$, the same result is obtained by modifying only the first component of the instruments vector?

(d) the matrix $\mathbf{M}^2$ gathers the coefficients which correspond to doubling every instrument, i.e., $\mathbf{M}^2\mathbf{x} = \mathbf{M}(2\mathbf{x})$?

9.12. In a financial market n financial assets are traded, with yields given by the vector

$$\mathbf{r} := \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix}.$$

A portfolio manager invests in the n assets the amounts

$$\mathbf{c} := \begin{bmatrix} c_1 & c_2 & \cdots & c_n \end{bmatrix}$$

Give the expression of the linear function "portfolio yield" $f(\mathbf{r})$ as a function of the single assets yields $\mathbf{r}$. What matrix $\mathbf{A}$ represents such a linear function?

10

Multivariable Differential Calculus

One-variable functions, such as $y = f(x)$, were illustrated in the previous chapters. However, they are inadequate for constructing a variety of models concerning Economics and Business Sciences. In fact we may find problems where many variables are involved, with various kinds of constraints, maybe even very complicated constraints, which cannot be described and analysed using one-dimensional methods. Therefore we need functions of several variables. From a conceptual point of view there is nothing new with respect to the one-dimensional case: we shall consider mappings from a set A into $\mathbb{R}$, associating with each element of A *only one* element of $\mathbb{R}$, but now A, which is the *domain* of f, is a subset of $\mathbb{R}^n$ instead of $\mathbb{R}$.

In this chapter, our purpose is to develop the fundamental features of multivariable differential calculus and optimization, restricting ourselves mostly to functions of two variables. In particular:

• We shall recall the main definitions, particularly the notions of (local and global) *extremum* and *extremum point*.

• We shall examine *quadratic functions* (or forms), paying particular attention to their sign.

• After a quick mention of *continuity,* we shall illustrate the concept of *partial derivative*.

• A brief analysis of level sets and of *implicit functions* will follow, with the important Dini's theorem.

• We shall then consider the second differential and *Taylor's formula*. In particular, the study of the sign of the second differential prepares the ground for *optimization* problems.

• We shall introduce first and second order conditions for *unconstrained optimization* problems.

• Finally we shall consider *constrained optimization* problems and discuss the important *method of Lagrange multipliers.*

10.1 Introduction

In Economics we often meet quantities depending on several variables. Let us consider a firm producing certain goods; the volume of production y depends on the quantities $x_1, x_2, \ldots, x_n$, representing capital, labour, raw materials... Such a dependence can be described by

$$y = f(x_1, x_2, \ldots, x_n),$$

where the number y is not determined by a single variable x, but by $x_1, x_2, \ldots, x_n$ jointly. In this case we say that f is a *function of several variables* and, more precisely, of n variables, also called the *arguments* of the function. A list of variables $x_1, x_2, \ldots, x_n$ can be thought of as an n-dimensional *vector*

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^{\mathrm{T}}$$

Therefore the function $y = f(x_1, x_2, \ldots, x_n)$ can be thought of as a *function of the vector* $\mathbf{x}$: $y = f(\mathbf{x})$. In this case, such a vector collects the quantities of each production factor used by the firm.

Functions of several variables are obviously more complex than functions of one variable; here we shall mostly refer to functions of two variables

$$z = f(x, y),$$

since many of their properties may be visualized geometrically. Moreover, many elementary economic theories are based on models with two-variable functions; therefore this basic scheme is really the case which is most frequently needed.

Examples

1.1. The volume V of a cylinder depends on the radius of its base r and on its height h, according to the formula

$$V = V(r, h) = \pi r^2 h,$$

which states that V is a function of the two variables r, h. Also the total surface A turns out to be a function of the two variables r, h:

$$A = A(r, h) = \pi r^2 + 2\pi r h = \pi r (r + 2h).$$

1.2. Many aggregate economic models assume that the gross national product Y of a country in a given year is connected with the availability of capital K and labour-power L by the relation

$$Y = g(K, L) = aK^{\alpha} L^{1-\alpha},$$

with $a > 0$ and $0 < \alpha < 1$. Such a relation is called a *Cobb-Douglas production function.*

10.1.1 Graph of two-variable functions

The graph of a function $f : A \subseteq \mathbb{R}^2 \to \mathbb{R}$ is generally represented by a surface in the three-dimensional space $\mathbb{R}^3$. More precisely:

$$\text{graph}\,(f) = \left\{(x, y, z) \in \mathbb{R}^3 : z = f\,(x, y)\right\}$$

Usually, we set three perpendicular axes in $\mathbb{R}^3$ in the directions of the unit vectors of the canonical basis. The points of A, which is the domain of f, lie on the "horizontal" plane, while the values assumed by f are represented on the "vertical" axis.

Examples

1.3. All points on the graph of the function

$$f(x, y) = 3$$

have the third coordinate equal to 3 and therefore the graph of f is a horizontal plane, parallel to the x, y-plane. The graph of any function of the type $f(x, y) = k$, where k is a real number, is a horizontal plane.

1.4. Let us consider the linear function

$$f(x, y) = -x - y.$$

Since $f(0, 0) = 0$, its graph "passes through" the origin and coincides with the plane in Figure 1. The graph of *any* linear function of two variables is a plane through the origin. The graph of the affine linear function

$$g(x, y) = 10 - x - y$$

is still a plane, more precisely it is the parallel plane obtained by an "upward translation" of the previous plane by 10 units (see Figure 1). The graph of *any* affine linear function of two variables is a plane.

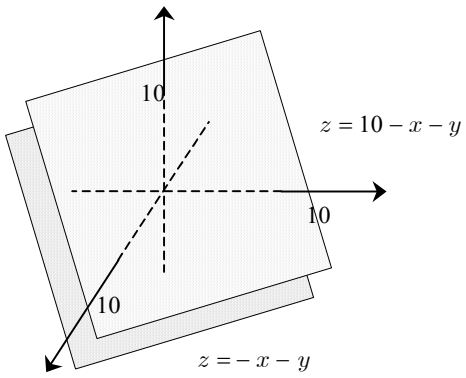


Figure 10.1. Parallel planes

1.5. Let us consider the Cobb-Douglas production function:

$$Y = 2K^{1/3}L^{2/3}, \quad K, L \geq 0. \tag{10.1}$$

In Figure 2 we show its three-dimensional graph.

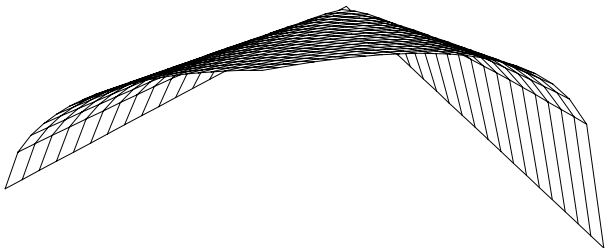


Figure 10.2. The Cobb-Douglas function $Y = 2K^{1/3}L^{2/3}$

The points of the horizontal plane represent the amounts of productive factors (K, L) used as input variables and, at each point of the vaulted graph, the height represents the level of production generated by such inputs. The vault is slightly asymmetric. This is due to the fact that the two productive factors appear in the function with different exponents. If we had represented the function $Y = 2K^{1/2}L^{1/2}$ we would have obtained a perfectly symmetric vault.

10.1.2 Level curves

The three-dimensional representation of the surface describing a function (like a production function) is not always convenient. It is possible to obtain a simple two-dimensional representation, which is very useful in many practical situations and which is similar to the one used for maps of mountain areas. We cut the surface horizontally with an appropriate number of parallel planes, so that we get a family of curves in the space. On each of these curves, the surface has a constant level. Then we represent the projections of such curves on a "geographical map", and they take the name of *level curves* (also: *contour lines* or *level contours*) of the surface. Plotting a level curve is equivalent to representing, in the x, y-plane, the set of all points such that

$$f(x, y) = c,$$

where c is one of the values assumed by the function.

Examples

1.6. Let us consider a linear function

$$z = f(x, y) = ax + by = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}.$$

The c-level curves in the x, y-plane are straight lines having the equation

$$ax + by = c$$

and they are perpendicular to the vector $\begin{bmatrix} a & b \end{bmatrix}$.

1.7. For the Cobb-Douglas function (10.1), the level curves are represented in Figure 3. The equation of the level curves is $2K^{1/3}L^{2/3} = c$, that is

$$K = \frac{c^3}{8L^2},$$

where $c > 0$ is the output level.

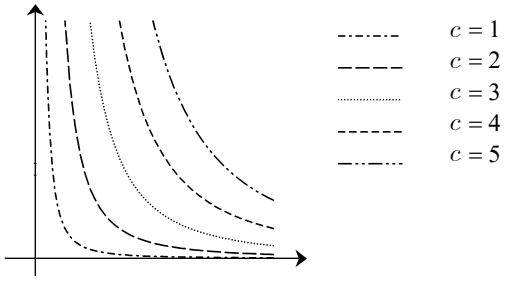


Figure 10.3. Level curves for a Cobb-Douglas function

1.8. The level curves for the function $f(x, y) = x^2 + y^2$, with $c > 0$, are circumferences with centre at the origin and radius $\sqrt{c}$. The level curve with $c = 0$ reduces to the point $(0, 0)$ and when $c < 0$ it is... the empty set.

10.2 Domain of a function

For a two-variable function, analytically represented by a formula $z = f(x, y)$, it is necessary to deal with the problem of determining its *natural domain*, that is the set of pairs (x, y) for which the function is defined. In the case of one-variable functions, in the presence of denominators, radicals of even index, logarithms... we have to write a system of conditions which must be jointly satisfied. For one-variable functions, it is usually easy to solve every single equation or inequality appearing as a condition and then take the intersection of the various sets of points solving them. For functions of two variables, we can still construct the solutions of the various equations or inequalities as subsets of the plane and then take their intersection.

However, the domains we might be dealing with can have a much more complicated structure. In order to better describe such domains, we start with some definitions.

Definition 2.1. *We call circular* ***neighbourhood*** *of a point* (x_0, y_0) *in* $\mathbb{R}^2$ *with radius* r *the circle* $B_r(x_0, y_0)$ *with centre at* (x_0, y_0) *and radius* r*, excluding the circumference.*

By means of the notion of circular neighbourhood (which will be simply called neighbourhood, from now on), we can refine the relation of belonging to a set and define some kinds of subsets of $\mathbb{R}^2$ which are particularly meaningful for future applications. Let us consider the set represented in Figure 4. We denoted the upper part of the contour, not belonging to the set, by a dashed line, and used a full line for the lower part of the contour, which belongs to the set.

We focus our attention on the four points $\mathbf{p}, \mathbf{a}_1, \mathbf{a}_2, \mathbf{q}$, and particularly on their relation with the set A.

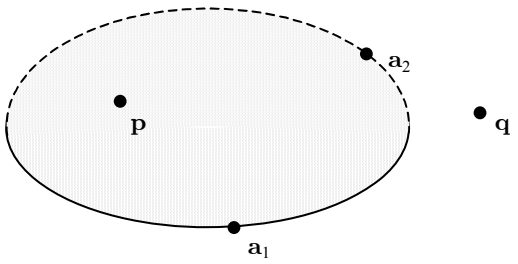


Figure 10.4. Interior, exterior and boundary points

The point $\mathbf{p}$ does not just belong to the set A, but it lies *well* inside the set: in fact, we can find a neighbourhood $B(\mathbf{p})$ of $\mathbf{p}$ which is fully contained in A. The point $\mathbf{q}$ not only does not belong to A, but it lies *well* outside it: there is a neighbourhood $B(\mathbf{q})$ of the point $\mathbf{q}$ which is fully contained in the complement of A. The two points $\mathbf{a}_1, \mathbf{a}_2$ are "halfway": the first lies on the lower part of the contour and therefore it belongs to the set, while the second lies on the other part of the contour and consequently it does not belong to the set. In both cases, we cannot say that they are "well inside or well outside" with respect to the set: any neighbourhood of these points (even if the radius is very small) is made up of some points belonging to A and some points belonging to the complement of A.

Definition 2.2. *Let* $A \subseteq R^2$*. We say that a point* $\mathbf{a}$ *is an* ***interior point*** *of* A*, if there is a neighbourhood* $B(\mathbf{a})$ *satisfying* $B(\mathbf{a}) \subseteq A$*; a point* $\mathbf{a}$ *is a* ***boundary point*** *of* A *if all its neighbourhoods contain both points of* A *and points of the complement of* A*.*

Now we define some kinds of sets, according to the "quality" of their points.

Definition 2.3. *If all the points of a set are interior, we say that the set is* ***open****. If a set contains all its boundary points, we say that it is* ***closed****.*

One can check that: *the complement of an open set is closed and the complement of a closed set is open.*

Examples

2.1. The domain of the function $f(x,y) = \sqrt{1-x^2-y^2}$ is the set of all points in the plane satisfying the inequality

$$1 - x^2 - y^2 \geq 0 \qquad \Longrightarrow \qquad x^2 + y^2 \leq 1,$$

that is the circle with centre at the origin and radius 1. This set is closed. Indeed, the boundary points for such a set are the points of the circumference of equation $x^2 + y^2 = 1$ and belong to the set itself.

The domain of the function $g(x,y) = 1/f(x,y)$, which corresponds to the set of all points in the plane satisfying the inequality $1 - x^2 - y^2 > 0$, that is to the circle without the circumference, is an open set.

Obviously there are sets which are neither open nor closed. $\mathbb{R}^2$ and the empty set $\emptyset$ are considered to be both open and closed; they are the only sets with this property.

Now we deal with those sets whose points have a finite distance from the origin. They are called bounded sets.

Definition 2.4. *A set A is said to be* ***bounded*** *if it is contained in some neighbourhood of the origin.*

Examples

2.2. The domains of the functions f and g in example 2.1 are examples of bounded sets.

2.3. The Cobb-Douglas production function $Y = aK^{1/2}L^{1/2}$ has the first quadrant $K, L \geq 0$ (including the non-negative half-axes) as its domain. It is a closed set, whose boundary is made up of the two non-negative half-axes. It is not a bounded set.

2.4. The function $z_1 = \ln(y - x)$ is defined only for $y > x$, that is above the bisector of the first and third quadrant; therefore it is necessary to exclude all points lying in the half-plane under such a bisector and the bisector itself. It is an open and unbounded set, illustrated in Figure 5 (a).

The bisector of the first and third quadrants is the boundary of the domain, but none of its points belongs to the domain.

2.5. The domain of the function $z_2 = \sqrt{y - x^2}$ is the set of all points in the (x,y)-plane such that $y \geq x^2$. These points lie above or on the parabola $y = x^2$, as shown in Figure 5 (b). The set is closed and unbounded.

2.6. The domain of the function $z = z_1 + z_2 = \ln(y - x) + \sqrt{y - x^2}$ is illustrated in Figure 5 (c). The set is neither open nor closed. It is not bounded.

10.3 Global and local extrema

We extend the concepts introduced for one-variable functions to functions of several variables. Some notions, such as boundedness, can be immediately generalised.

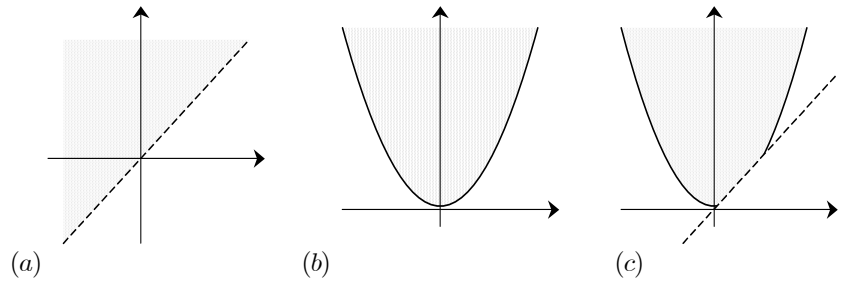


Figure 10.5. The domains of the functions in examples 2.4, 2.5 and 2.6

Definition 3.1. *A function $f : A \subseteq \mathbb{R}^2 \to \mathbb{R}$ is said to be* ***bounded from above*** *(**below**), if there is a number K such that, for every $(x, y) \in A$, it turns out to be $f(x, y) \leq K$ ($f(x, y) \geq K$); if f is bounded both from below and from above, we say that f is* ***bounded***.

The geometric meaning is clear. If f is bounded from above by K, it means that its graph never goes above the horizontal plane of equation $z = K$. Similarly, saying that a function f is bounded from below means that its graph never goes below the plane of equation $z = K$.

Now we illustrate the notions of extremum and extremum point, by means of two simple examples.

Example 3.1. Let us consider the function of two variables (Figure 6 (a))

$$f(x, y) = 1 - x^2 - y^2.$$

We note that

$$f(0, 0) = 1 - 0^2 - 0^2 = 1$$

so that at the point $(x, y) = (0, 0)$ the function takes the value 1. At any other point $(x, y) \in \mathbb{R}^2$, the value of the function is less then 1. Thus

$$f(0, 0) > f(x, y) \qquad \text{for every } (x, y) \neq (0, 0).$$

We say that $(0, 0)$ is a *strict global* (or *strong* global) maximum point and that 1 is the maximum for f.

Of course, it may happen that a function does not have a strict global maximum point (x^*, y^*) because there are other points $\mathbf{x}$ where f assumes the same value. In such cases, we refer to a *weak global* maximum (or simply a global maximum).

Example 3.2. Let us consider the function

$$z = f(x, y) = 2 - x^2,$$

which is constant with respect to y. Its graph is generated by shifting the parabola of equation $z = 2 - x^2$, which we can easily draw in the x, z-plane, along the y-axis

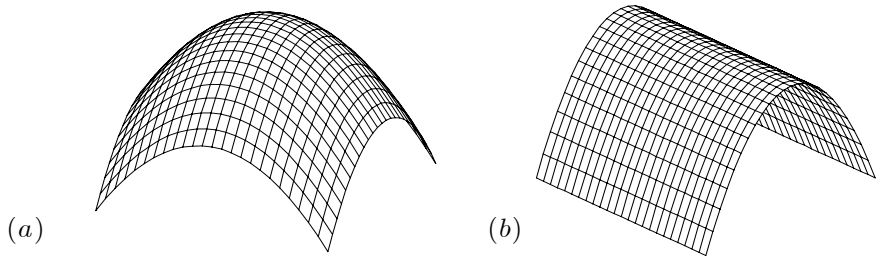


Figure 10.6. Strict maximum point and weak maximum points

(Figure 6 (b)). At every point of the y-axis the value of the function is 2, elsewhere it is less than 2. The value 2 is a global maximum but, being assumed at infinitely many points, it is a *weak maximum.*

Of course, being unbounded from below, neither of the functions in the two preceding examples takes a minimum value.

Definition 3.2. *Let $f : A \subseteq \mathbb{R}^2 \to \mathbb{R}$. We say that $(x^*, y^*) \in A$ is a **global maximum (minimum) point** and that $f(x^*, y^*)$ is the **global maximum** (**minimum**) for f on A, if, for every $(x, y) \in A$, $(x, y) \neq (x^*, y^*)$, it turns out that*

$$f(x^*, y^*) \geq f(x, y) \qquad (f(x^*, y^*) \leq f(x, y)). \tag{10.2}$$

If in (10.2) we have strict inequalities, we call the point a *strict global maximum* (*minimum*) *point.* As in the case of one-variable functions, we call *extremum points* the maximum and minimum points and *extrema* the maximum and minimum values assumed by the function. We recall also the respective local notions, for which we use the notion of *neighbourhood,* already introduced at page 316.

Definition 3.3. *Let $f : A \subseteq \mathbb{R}^2 \to \mathbb{R}$.We say that $(x^*, y^*) \in A$ is a **local maximum** (**minimum**) **point** and that $f(x^*, y^*)$ is a **local maximum** (**minimum**) for f on A, if there exists a neighbourhood U of (x^*, y^*) such that, for every $(x, y) \in A \cap U$, $(x, y) \neq (x^*, y^*)$,*

$$f(x^*, y^*) \geq f(x, y) \qquad (f(x^*, y^*) \leq f(x, y)). \tag{10.3}$$

Again, if in (10.3) the inequalities are strict, we call it a *strict local maximum* (*minimum*) *point.* As we already noted for one-variable functions, *every global maximum or minimum point is also a local maximum or minimum point.*

10.3.1 Concave and convex functions

We recall the notion of *convex* set. We shall discuss it in $\mathbb{R}^n$, even if we shall usually limit ourselves to dimensions 2 and 3. In $\mathbb{R}^n$ the *segment connecting two points* $\mathbf{p}$ and $\mathbf{q}$ is the set of all *convex* linear combinations of the vectors $\mathbf{p}$ and $\mathbf{q}$, that is the set of all points of the type

$$\mathbf{r}_t = (1 - t)\,\mathbf{p} + t\mathbf{q}$$

when t varies in the interval $[0,1]$. A set $E \subseteq \mathbb{R}^n$ is *convex* if, given any pair of points in it, the segment connecting the two points is completely contained in the set itself.

Now for the sake of simplicity let us assume $f : A \subseteq \mathbb{R}^2 \to \mathbb{R}$, and let us consider the *epigraph* of f, defined by

$$\mathrm{Epi}\,(f) = \left\{(x,y,z) \in \mathbb{R}^3 :\ z \geq f\,(x,y)\,,\ (x,y) \in A\right\},$$

that is the set of all points (x,y,z) in $\mathbb{R}^3$ which lie above or on the graph of the function, while the pair (x,y) varies in the domain A.

Definition 3.4. *$f : A \subseteq \mathbb{R}^2 \mapsto \mathbb{R}$ is said to be **convex** if its epigraph is a convex set in $\mathbb{R}^3$; f is **concave** if $-f$ is convex.*

We note that the definition of a convex or concave function automatically implies that the domain A must be convex. As in the one-dimensional case, also for two-variable functions the convexity of the epigraph is equivalent to requiring that every segment connecting any two points on the graph of f has to lie above or at least not below the graph of f. This condition may be analytically translated by writing that, for every pair of points $\mathbf{p} = (x_1, y_1)$ and $\mathbf{q} = (x_2, y_2)$ in A and for every $t \in [0,1]$, the following inequality must hold:

$$f\,[(1-t)\,\mathbf{p} + t\mathbf{q}] \leq (1-t)\,f\,(\mathbf{p}) + tf\,(\mathbf{q})\,. \tag{10.4}$$

If f is concave, relation (10.4) holds with $\geq$ instead of $\leq$. If for $t \in (0,1)$ the *strict* sign holds in formula (10.4), then the function is said to be *strictly convex.*

Typical *strictly convex* (*concave*) functions in the whole $\mathbb{R}^n$ are the positive (negative) definite quadratic forms, which will be introduced in the next section. A positive (negative) semi-definite quadratic form is convex (concave), but not strictly convex (concave). Convex/concave functions satisfy a property which is important for their optimization, known as the *local/global theorem*:

Theorem 3.1. *Let $f : A \subseteq \mathbb{R}^2 \to \mathbb{R}$ be convex (concave). Every local minimum (maximum) point is also a global minimum (maximum) point.*

10.4 Quadratic forms

In optimization theory, a particularly important class of functions is the class of all second-degree homogeneous polynomials, often called *quadratic forms.* A generic quadratic form in two variables is given by

$$f\,(x_1, x_2) = m_{11}x_1^2 + 2m_{12}x_1x_2 + m_{22}x_2^2. \tag{10.5}$$

Let us examine three typical examples, whose graphs are drawn in Figures 7 (a), (b), (c): they represent, respectively, the functions

$$f\,(x_1, x_2) = x_1^2 + x_2^2, \quad g\,(x_1, x_2) = x_1^2, \quad h\,(x_1, x_2) = x_1^2 - x_2^2.$$

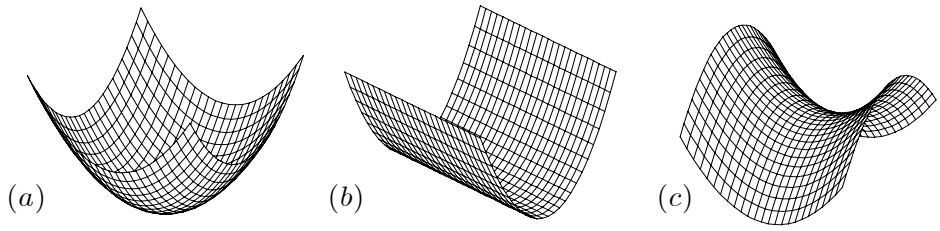


Figure 10.7. Quadratic forms: positive definite, positive semi-definite and indefinite

For the first function the point $(0,0)$ is a strict global minimum point, for the second it is a weak global minimum point, while for the third it is not an extremum point; it is anyway a particular point and it has an appropriate name: *saddle point*.

The different behaviours depend on the sign assumed by the quadratic form as $\mathbf{x} = (x_1, x_2)$ varies. In fact, we always have $f(0,0) = 0$, and therefore $\mathbf{0} = (0,0)$ is a maximum point if for every pair (x_1, x_2) it turns out that $f(x_1, x_2) \leq 0$, while it is a minimum point if for every pair (x_1, x_2) we have $f(x_1, x_2) \geq 0$. If there are some points where $f(x_1, x_2)$ is positive and other points where it is negative, then $(0,0)$ is a saddle point. The following table gathers all possible cases and the respective names for the three disjoint classes of quadratic forms.

quadratic form	**sign**	**0 is a**
positive / negative definite	> 0 / < 0 if $\mathbf{x} \neq \mathbf{0}$	strict global minimum point / strict global maximum point
positive / negative semi-definite	≥ 0 / ≤ 0 and (10.5) is null for some $\mathbf{x} \neq \mathbf{0}$	weak global minimum point / weak global maximum point
indefinite	changes with $\mathbf{x}$	saddle point

With the new terminology, the three quadratic forms f, g, h are, respectively, *positive definite, positive semi-definite* and *indefinite*. We want to develop a handy method in order to identify the nature of a general quadratic form. Let us begin by noting that the form (10.5) can be associated with the second order symmetric matrix

$$\mathbf{M} = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix},$$

with $m_{21} = m_{12}$ and that, setting $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, we can write

$$f(x_1, x_2) = m_{11}x_1^2 + 2m_{12}x_1x_2 + m_{22}x_2^2 = \mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x}.$$

For example, for the above functions f, g and h we have

$$\mathbf{M}_f = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{M}_g = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{M}_h = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

In practice, a quadratic form may be identified with the associated matrix, so that we can transfer the classification of the table to the matrix itself. Therefore we can talk directly of a *definite, semi-definite, indefinite matrix* instead of the associated form.

Now we study the sign of $f(x_1, x_2) = \mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x}$. If m_{11} and m_{22} are null, the quadratic form (10.5) reduces to $2m_{12}x_1x_2$, which is clearly indefinite because the two factors x_1, x_2 may have any sign as the pair (x_1, x_2) varies.

Let us suppose, therefore, that $m_{11} \neq 0$: the other case is perfectly analogous. The expression on the right hand side of (10.5) may be re-written, if we suppose that $x_2 \neq 0$[1] and we factor out x_2^2:

$$\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x} = x_2^2 \left[\left(\frac{x_1}{x_2} \right)^2 m_{11} + 2 \left(\frac{x_1}{x_2} \right) m_{12} + m_{22} \right].$$

Setting $z = x_1/x_2$ we get the representation:

$$\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x} = x_2^2 \left(m_{11}z^2 + 2m_{12}z + m_{22} \right).$$

It is clear that the factor x_2^2 (> 0) has no influence on the sign of the product, therefore we are only interested in the sign of the second degree trinomial

$$T(z) = m_{11}z^2 + 2m_{12}z + m_{22}.$$

If we want $\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x}$ to be positive for any choice of $\mathbf{x}$, $T(z)$ has to be positive for every[2] $z \neq 0$. If we want it to be negative, the trinomial has to be negative. In order to have an always positive or an always negative $T(z)$, for all values of $z \neq 0$, it is necessary and sufficient that:

— the coefficient m_{11} of z^2 is positive in the first case, negative in the second,

— the discriminant of the trinomial is negative.

Since the discriminant of $T(z)$ is $4m_{12}^2 - 4m_{11}m_{22} = 4\left(m_{12}^2 - m_{11}m_{22}\right)$, the second condition is equivalent to[3]

$$\det \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} = m_{11}m_{22} - m_{12}^2 > 0.$$

Therefore, in this case the quadratic form (10.5) is positive definite if $m_{11} > 0$, negative definite if $m_{11} < 0$.

In order that $T(z)$ change its sign as z varies, it is necessary and sufficient that the discriminant of the trinomial is positive, that is

$$\det \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} = m_{11}m_{22} - m_{12}^2 < 0.$$

---

[1] Since $(x_1, x_2) \neq (0, 0)$, we can always suppose that $x_2 \neq 0$. The argument is similar if we suppose $x_1 \neq 0$.

[2] z is the ratio between the two components of $\mathbf{x}$, therefore it can assume any real value.

[3] We note that this condition can be satisfied only in the case where m_{11} and m_{22} have the same sign (and are not null).

This also includes the case $(m_{11}, m_{22}) = (0, 0)$ we considered at the beginning: the form (10.5) is therefore indefinite.

We still have to study the case $m_{11}m_{22} - m_{12}^2 = 0$, for which the trinomial becomes

$$m_{11}\left(z + \frac{m_{12}}{m_{11}}\right)^2.$$

We see that $T(z)$ is null for $z = -m_{12}/m_{11}$; in all other cases it has the same sign as m_{11}. The form (10.5) is therefore semi-definite with the sign of m_{11}.

If $m_{11}m_{22} - m_{12}^2 = 0$ and $m_{11} = 0$, a similar argument shows that the form (10.5) is semi-definite with the sign of m_{22}; while if $m_{11}m_{22} - m_{12}^2 = 0$ and m_{11}, m_{22} are both different from zero, they have the same sign.

Therefore, when $m_{11}m_{22} - m_{12}^2 = 0$ we can always refer to the sign of the quantity $m_{11} + m_{22}$: the form (10.5) is semi-definite with the sign of $m_{11} + m_{22}$.

We summarize the results we have just achieved in another table, which completes the previous one.

condition on the coefficients	**quadratic form**
$\det(\mathbf{M}) > 0$, $\begin{matrix} m_{11} > 0 \\ m_{11} < 0 \end{matrix}$	$\begin{matrix} \text{positive} \\ \text{negative} \end{matrix}$ definite
$\det(\mathbf{M}) = 0$, $\begin{matrix} m_{11} + m_{22} > 0 \\ m_{11} + m_{22} < 0 \end{matrix}$	$\begin{matrix} \text{positive} \\ \text{negative} \end{matrix}$ semi-definite
$\det(\mathbf{M}) < 0$	indefinite

Example 4.1. Let

$$q_1(x_1, x_2) = 3x_1^2 + x_1x_2 - x_2^2, \quad q_2(x_1, x_2) = -2x_1^2 + 6x_1x_2 - 5x_2^2.$$

We have:

$$\begin{aligned} \mathbf{M}_{q_1} &= \begin{bmatrix} 3 & 1/2 \\ 1/2 & -1 \end{bmatrix}, \ \det(\mathbf{M}_{q_1}) = -3 - \frac{1}{4} < 0 \\ \mathbf{M}_{q_2} &= \begin{bmatrix} -2 & 3 \\ 3 & -5 \end{bmatrix}, \ \det(\mathbf{M}_{q_2}) = 10 - 9 > 0, \ m_{11} = -2 < 0. \end{aligned}$$

The conclusions are: q_1 is indefinite, q_2 is negative definite.

Quadratic forms in n variables

Of course, we can construct quadratic forms, i. e. second degree homogeneous polynomials, in any number of variables. A general quadratic form with n variables $x_1, \ldots, x_n$ can be written as

$$\begin{aligned} \sum_{i,j=1}^{n} m_{ij}x_ix_j = {} & m_{11}x_1^2 + m_{12}x_1x_2 + \cdots + m_{1n}x_1x_n + \\ & + m_{21}x_2x_1 + m_{22}x_2^2 + \cdots + m_{2n}x_2x_n + \\ & + \cdots + \\ & + m_{n1}x_nx_1 + m_{n2}x_nx_2 + \cdots + m_{nn}x_n^2, \end{aligned}$$

or also as $\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x}$, with $\mathbf{x}^{\mathrm{T}} = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}$ and $\mathbf{M} = [m_{rs}]$ $(r, s = 1, \ldots, n)$. We can always suppose that $\mathbf{M}$ is symmetric.

Example 4.2. The polynomial

$$3x_1^2 + 2x_1x_2 + 7x_2^2 - 4x_2x_3 + 5x_3^2 = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 3 & 1 & 0 \\ 1 & 7 & -2 \\ 0 & -2 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

is a three-variable quadratic form.

As for two-variable quadratic forms, it is important to understand whether $\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x}$, which is null for $\mathbf{x} = \mathbf{0}$, changes its sign or not when $\mathbf{x}$ varies. The classification of quadratic forms with n variables (or associated matrices) according to their sign is unchanged with respect to the two-dimensional case. We can take the table on page 321 and consider $\mathbf{x}$ as an n-dimensional vector; everything remains identical.

Again we want to find an effective method for classifying a quadratic form. A new reading of the two-dimensional case might be useful. The second table tells us that if the two numbers

$$m_{11} = \det[m_{11}] \quad \text{and} \quad m_{11}m_{22} - m_{12}^2 = \det(\mathbf{M}) \tag{10.6}$$

are positive then the form is positive definite, while if the first is negative and the second is positive the form turns out to be negative definite.

The two numbers in (10.6) are nothing but the North-West principal minors[4] of the matrix $\mathbf{M}$. For quadratic forms with three variables, the matrix $\mathbf{M}$ has order three and it has three North-West principal minors:

$$\det[m_{11}]\ , \quad \det\begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \quad \text{and} \quad \det\begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix}.$$

In this case $\mathbf{M}$ *is positive definite if and only if the signs of its NW principal minors are positive* $(+,+,+)$*, while it is negative definite if and only if those signs are alternate starting with a negative sign* $(-,+,-)$.

The quadratic function in example 4.2

$$f(x_1, x_2, x_3) = 3x_1^2 + 2x_1x_2 + 7x_2^2 - 4x_2x_3 + 5x_3^2$$

is positive definite and, therefore, it has a strong global minimum point at $(0, 0, 0)$, because

$$3 > 0, \quad \det\begin{bmatrix} 3 & 1 \\ 1 & 7 \end{bmatrix} = 20 > 0, \quad \begin{bmatrix} 3 & 1 & 0 \\ 1 & 7 & -2 \\ 0 & -2 & 5 \end{bmatrix} = 88 > 0.$$

[4] A North-West principal minor of order k is the determinant of the submatrix contructed with the first k rows and the first k columns of $\mathbf{M}$.

If $\det(\mathbf{M}) \neq 0$ and the chain of signs is different from $+,+,+$ or $-,+,-$, the form $\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x}$ certainly changes its sign depending on $\mathbf{x}$; in this case it is *indefinite*.

If $\det(\mathbf{M}) = 0$ the form $\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x}$ is zero also for some non-zero value of $\mathbf{x}$; it might change its sign, but also be ≥ 0 or ≤ 0 for every $\mathbf{x}$; a distinction is more complicated in these cases.

The results in the three-dimensional case can be extended to the general case of n variables: this is the subject of our next theorem.

Theorem 4.1. *A symmetric square matrix* $\mathbf{M}$ *is positive definite if and only if the signs of its North-West principal minors*

$$m_{11},\ \det\begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix},\ \det\begin{bmatrix} m_{11} & \cdots & m_{1k} \\ \vdots & \ddots & \vdots \\ m_{k1} & \cdots & m_{kk} \end{bmatrix},\ \ldots,\ \det(\mathbf{M})$$

are all positive; it is negative definite if and only if such signs are alternate, starting from $m_{11} < 0$.

10.5 Continuity

The intuitive idea of continuity for functions of several variables is the same we have already seen in the scalar case: a small variation in the input produces a small variation in the output. Let us consider a function $f : A \subseteq \mathbb{R}^2 \to \mathbb{R}$ and a point $\mathbf{p}_0 = (x_0, y_0) \in A$; f is *continuous* at $\mathbf{p}_0$, if, for every $\mathbf{p} = (x, y) \in A$ having a "small distance" from $\mathbf{p}_0$, the value $f(\mathbf{p})$ has a "small distance" from $f(\mathbf{p}_0)$. That is, if

$$|\mathbf{p} - \mathbf{p}_0| \to 0 \quad \Longrightarrow \quad |f(\mathbf{p}) - f(\mathbf{p}_0)| \to 0.$$

If a function is continuous at every point of a set it is said to be *continuous* on such a set. As in the one-dimensional case, elementary functions (exponentials, power functions, logarithms, trigonometric functions) are continuous, as well as all functions obtained by operations (such as the usual algebraic operations, or the composition operation) on continuous functions. In particular all linear functions $f(x, y) = ax + by$, all quadratic functions $f(x, y) = ax^2 + bxy + cy^2$ and all functions like

$$f(x, y) = 3x^4y^3 - 15x^2y^7, \qquad g(x, y) = x \ln y + ye^x$$

are continuous on their domain.

In the one-dimensional case, continuous functions defined on closed and bounded intervals $[a, b]$ have global extrema. This is the simplest version of Weierstrass's theorem. In the general version, the interval $[a, b]$ can be substituted by a closed and bounded set. We can now state its two-dimensional analogous.

Theorem 5.1. *If* $f : A \subset \mathbb{R}^2 \to \mathbb{R}$ *is continuous and* A *is closed and bounded, then* f *attains a global maximum and a global minimum on* A; *in other words, there are (at least) two points* $\mathbf{p}_1, \mathbf{p}_2 \in A$ *such that*

$$\min_A f = f(\mathbf{p}_1) \leq f(\mathbf{p}) \leq f(\mathbf{p}_2) = \max_A f, \quad \textit{for every } \mathbf{p} \in A.$$

10.6 Partial derivatives

We now extend differential calculus to functions of two variables, $f : A \subseteq \mathbb{R}^2 \to \mathbb{R}$, where A is, for now, an *open* set, and therefore its points are all interior points. This fact allows us to slightly move away from any of them in every direction, while still remaining inside A. The first notion we want to generalize is that of derivative at a point $(x_0, y_0) \in A$. Since there are two variables, the idea is to start from (x_0, y_0) and separately increase the two variables x and y in the directions of the axes. This way we get two functions of one variable, given by

$$x \to f(x, y_0) \qquad \text{and} \qquad y \to f(x_0, y), \tag{10.7}$$

for which the notion of derivative is well-known. Thus we arrive at the concept of partial derivative, which is formalized in the following definition.

Definition 6.1. *The **partial derivatives** of f at (x_0, y_0), with respect to the variables x and y, respectively, are defined by the two following limits, if they exist and are finite*:

$$\begin{aligned} f'_x(x_0, y_0) &= \lim_{h \to 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h}, \\ f'_y(x_0, y_0) &= \lim_{k \to 0} \frac{f(x_0, y_0 + k) - f(x_0, y_0)}{k}. \end{aligned}$$

As we can see, in the case of the derivative with respect to x we consider x as a variable and keep y as a constant. The opposite holds in the case of the derivative with respect to y. Other commonly used symbols for partial derivatives are:

$$\frac{\partial f}{\partial x}(x_0, y_0), \quad \frac{\partial f}{\partial y}(x_0, y_0) \quad \text{or} \quad D_x f(x_0, y_0), \quad D_y f(x_0, y_0),$$

recalling the symbols introduced by Cauchy.

Once the partial derivatives at a point are defined, we go on to *partial derivative functions.* If $f'_x(x, y)$ and $f'_y(x, y)$ exist for all $(x, y) \in A$, the functions

$$f'_x : A \subseteq \mathbb{R}^2 \to \mathbb{R} \quad \text{and} \quad f'_y : A \subseteq \mathbb{R}^2 \to \mathbb{R}$$

are defined; they associate with every $(x, y) \in A$ the values $f'_x(x, y)$ and $f'_y(x, y)$, respectively.

With the partial derivatives of a function f we can construct a vector (usually written as a row vector) which is very important and deserves a special name.

Definition 6.2. *Let $f : A \subseteq \mathbb{R}^2 \to \mathbb{R}$ and $(x, y) \in A$.The vector*

$$\nabla f(x, y) = \begin{bmatrix} f'_x(x, y) & f'_y(x, y) \end{bmatrix}$$

*is called the **gradient** of f at (x, y) and is denoted by one of the symbols $f'(x, y)$ or $\nabla f(x, y)$.*

In most common cases, partial derivatives are computed by means of the usual rules seen in section 6.4.

Examples

6.1. Let $f(x,y) = \ln \dfrac{x+2y}{x-3y}$; by applying the chain rule and the quotient rule, we get:

$$\begin{aligned} f'_x(x,y) &= \frac{x-3y}{x+2y} \cdot \frac{x-3y-(x+2y)}{(x-3y)^2} = \frac{-5y}{(x+2y)(x-3y)}, \\ f'_y(x,y) &= \frac{x-3y}{x+2y} \cdot \frac{2(x-3y)+3(x+2y)}{(x-3y)^2} = \frac{5x}{(x+2y)(x-3y)}. \end{aligned}$$

6.2. Let $g(x,y) = x^y$; the x-derivative is the derivative of a power function, the y-derivative is the derivative of an exponential. Thus we get

$$g'_x(x,y) = yx^{y-1}, \qquad g'_y(x,y) = x^y \ln x.$$

• *Geometric meaning.* The geometric meaning of the partial derivatives is analogous to the one-dimensional case. The graph of a function f of two variables is a surface in the three-dimensional space. Intersecting such a surface with the planes $y = y_0$ or $x = x_0$, perpendicular to the x,y-plane, we get the graph of the elementary functions (10.7). Let $z_0 = f(x_0,y_0)$. The derivative $f'_x(x_0,y_0)$ is the slope of the tangent to the graph of $x \longmapsto f(x,y_0)$ at the point (x_0,y_0,z_0) (Figure 8), while $f'_y(x_0,y_0)$ is the slope of the tangent line to the graph of $y \longmapsto f(x_0,y)$ at the point (x_0,y_0,z_0).

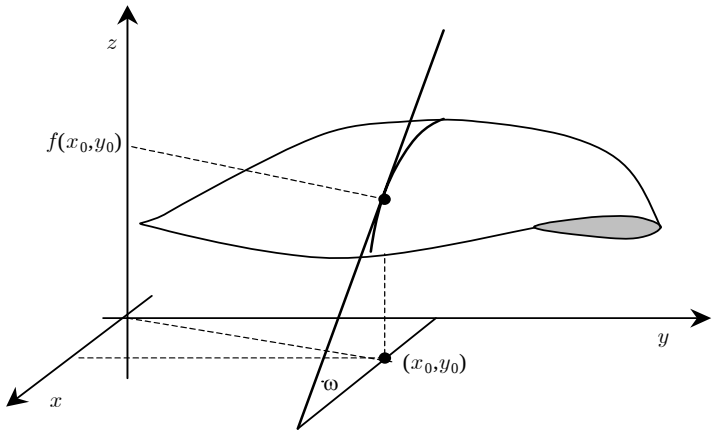


Figure 10.8. Geometric meaning of the x-derivative: $f'_x(x_0,y_0) = \tan\omega$

10.7 Differentiability and tangent plane

For one-variable functions, the existence of the derivative at a point is equivalent to the existence of the tangent line to the graph at the corresponding point. For

functions of two variables, whose graph is a surface, the concept of tangent line is naturally replaced by the concept of *tangent plane.* This plane should well approximate the graph of the surface in a neighbourhood of the tangency point. Now, it happens that the mere existence of the two partial derivatives *does not guarantee* at all that such a plane exists. The appropriate notion is the *differentiability,* which plays, in the case of two variables as well as in the general case of n variables, a role which is much more important than the existence of partial derivatives. This is a remarkable difference with respect to the one-dimensional case, where the existence of the derivative and differentiability imply each other. Let us introduce the definition of differentiability.

Definition 7.1 *Let $f : A \subseteq \mathbb{R}^2 \to \mathbb{R}$, and $(x_0, y_0) \in A$. We say that f is* ***differentiable*** *at (x_0, y_0) if there is a vector $\mathbf{m} = (m_1, m_2) \in \mathbb{R}^2$, such that, for every vector of increments $(h, k) \in \mathbb{R}^2$ satisfying $(x_0 + h, y_0 + k) \in A$, the increment of f may be written, as $(h, k) \to (0, 0)$,*

$$f(x_0 + h, y_0 + k) - f(x_0, y_0) = m_1 h + m_2 k + o\left(\sqrt{h^2 + k^2}\right). \tag{10.8}$$

We note that a differentiable function is always continuous. Indeed, as one can easily see from (10.8), if (h, k) tends to $(0, 0)$, the right hand side tends to 0. Thus $f(x_0 + h, y_0 + k)$ tends to $f(x_0, y_0)$; in other words, the function f is continuous at (x_0, y_0).

The *linear* function[5] of (h, k)

$$(h, k) \mapsto m_1 h + m_2 k$$

takes the name of *(first) differential* of f at the point (x_0, y_0) and is denoted by the symbol $\mathrm{d}f(x_0, y_0)$. We can prove that the vector $\mathbf{m}$ *coincides* with the gradient[6] of f at the point (x_0, y_0), so that the product $m_1 h + m_2 k$ becomes

$$f'_x(x_0, y_0)\, h + f'_y(x_0, y_0)\, k.$$

---

[5] If we see vectors as particular cases of matrices, coherently with the representation theorem, we should write

$$\begin{bmatrix} h \\ k \end{bmatrix} \mapsto \begin{bmatrix} m_1 & m_2 \end{bmatrix} \begin{bmatrix} h \\ k \end{bmatrix}.$$

[6] Let $k = 0$ in (10.8): we get

$$f(x_0 + h, y_0) - f(x_0, y_0) = m_1 h + o(h).$$

Dividing by h and taking the limit for $h \to 0$ we get

$$\frac{\partial f}{\partial x}(x_0, y_0) = \lim_{h \to 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h} = m_1.$$

Letting $h = 0$ in (10.8) and proceeding in the same way, we obtain

$$\frac{\partial f}{\partial y}(x_0, y_0) = m_2.$$

Introducing the differentials $\mathrm{d}x$ and $\mathrm{d}y$ (which can intuitively be associated with "infinitesimal" increments of the independent variables), we can write the differential in the more significant form

$$\mathrm{d}f(x_0, y_0) = f'_x(x_0, y_0)\,\mathrm{d}x + f'_y(x_0, y_0)\,\mathrm{d}y \tag{10.9}$$

Now let us define the tangent plane.

If f is differentiable we have, setting $h = x - x_0$, $k = y - y_0$:

$$\begin{aligned} f(x,y) &= f(x_0,y_0) + f'_x(x_0,y_0)(x-x_0) + f'_y(x_0,y_0)(y-y_0) + \\ &\quad + o\left(\sqrt{(x-x_0)^2 + (y-y_0)^2}\right) \quad \text{as } (x,y) \to (x_0,y_0). \end{aligned} \tag{10.10}$$

Letting $z_0 = f(x_0, y_0)$, formula (10.10) reveals that the graph of $z = f(x, y)$ near to the point (x_0, y_0, z_0) is well approximated by the (affine) linear function

$$\boxed{z = z_0 + f'_x(x_0, y_0)(x - x_0) + f'_y(x_0, y_0)(y - y_0).} \tag{10.11}$$

Formula (10.11) *defines the* **tangent plane** to the surface of equation $z = f(x, y)$ at the point (x_0, y_0, z_0).

Example 7.1. The equation of the tangent plane to the graph of the two-variable function

$$z = f(x, y) = 10 - x^2 - y^2$$

at the point $(1, 1)$ is

$$\begin{aligned} z &= f(1,1) + f'_x(1,1)(x-1) - f'_y(1,1)(y-1) = \\ &= 8 - 2(x-1) - 2(y-1) = -2x - 2y + 12. \end{aligned}$$

• *Level curves.* Let $z = g(x, y)$ be the equation of a surface in the three-dimensional space with g differentiable at a point (x_0, y_0). We know that, setting $z_0 = g(x_0, y_0)$, the equation of the tangent plane at (x_0, y_0, z_0) is given by formula (10.11). Now let us cut the graph of g with the horizontal plane $z = z_0$ and consider, in the x, y-plane, the corresponding level curve of equation

$$g(x, y) = z_0. \tag{10.12}$$

If we now cut the graph of the tangent plane (10.11) with the plane $z = z_0$, the corresponding level curve will be the straight line of equation

$$g'_x(x_0, y_0)(x - x_0) + g'_y(x_0, y_0)(y - y_0) = 0 \tag{10.13}$$

obtained from formula (10.11) by putting $z = z_0$. Of course, if we want formula (10.13) to be a true straight line, it is necessary to require that $\nabla g(x_0, y_0) \neq \mathbf{0}$.

This straight line turns out to be tangent to the level curve (10.12) at the point (x_0, y_0) and may be re-written as

$$\begin{bmatrix} g'_x(x_0,y_0) & g'_y(x_0,y_0) \end{bmatrix} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix} = 0. \tag{10.14}$$

The meaning of (10.14) is rather interesting. It states that the scalar product between $\nabla g\,(x_0, y_0)$ and the vector $\begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix}$, which is parallel to the straight line (10.13), is null. The two vectors are therefore orthogonal.

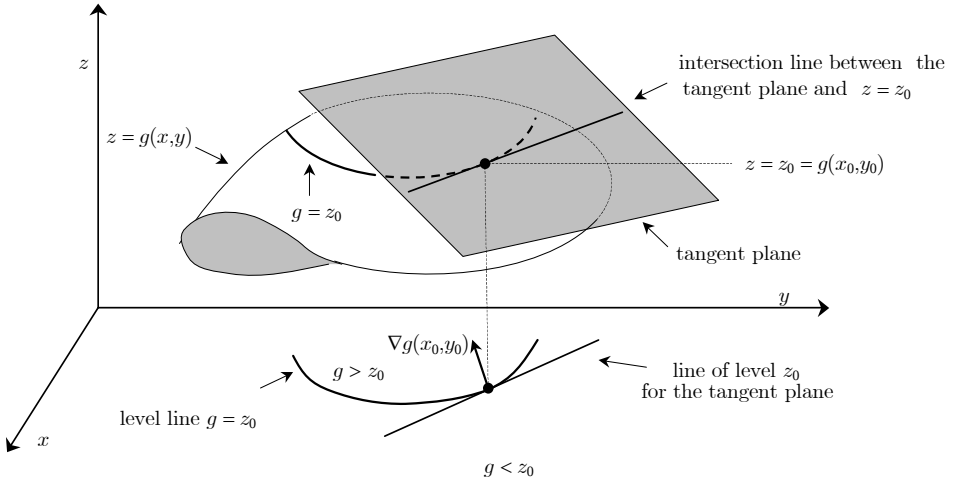


Figure 10.9.

Proposition 7.1. *At every point* (x_0, y_0) *such that* $\nabla g\,(x_0, y_0) \neq \mathbf{0}$, $\nabla g\,(x_0, y_0)$ *is orthogonal to the level curve passing through that point.*

Everything is shown in Figure 9, where we have also indicated the direction towards which the gradient is pointing, if it is not null: it is always directed towards the region where $g > z_0$.

The mere existence of partial derivatives does not guarantee differentiability. However, if at least one of the partial derivatives is continuous, then the function is differentiable. More precisely, we can state the following

Theorem 7.2 (Sufficient condition for differentiability). *Let* $f : A \subseteq \mathbb{R}^2 \to \mathbb{R}$ *and* $(x_0, y_0) \in A$. *If*:
(*i*) f'_x *and* f'_y *exist in a neighbourhood of* (x_0, y_0) *and*
(*ii*) *at least one of them is continuous at the point* (x_0, y_0),
then f *is differentiable at* (x_0, y_0).

According to theorem 7.2, almost all elementary functions are differentiable.

For example, power functions with integer exponent, exponentials, logarithmic and trigonometric functions are differentiable at the interior points of their domain. Furthermore, we can prove that the sum, product, quotient (where the denominator does not vanish) and composition of these functions are differentiable.

• *Directional derivatives.* The partial derivatives at a point (x_0, y_0) are nothing else but the derivatives of the restriction of the function f to straight lines parallel to the two coordinate axes. We can also restrict f to any straight line through (x_0, y_0) and compute the derivative of such a (one-variable) function. In this way we are led to the notion of a *directional derivative.*

Let $f : A \subseteq \mathbb{R}^2 \to \mathbb{R}$, (x_0, y_0) be an interior point for A and $\mathbf{v}$ be a unit vector

$$\mathbf{v} = (\cos\alpha, \sin\alpha).$$

We define the *directional derivative* of f at the point (x_0, y_0) in the direction of the unit vector $\mathbf{v}$, as the following limit, if it exists and is finite:

$$\lim_{t\to 0} \frac{f(x_0 + t\cos\alpha, y_0 + t\sin\alpha) - f(x_0, y_0)}{t}.$$

The derivative of f at the point (x_0, y_0) in the direction of the unit vector $\mathbf{v}$ is denoted by the symbol $D_{\mathbf{v}} f(x_0, y_0)$. More simply, given

$$g(t) = f(x_0 + t\cos\alpha, y_0 + t\sin\alpha),$$

we have

$$D_{\mathbf{v}} f(x_0, y_0) = g'(0).$$

If the function f is differentiable at the point (x_0, y_0), the derivatives along every direction can be expressed in terms of the partial derivatives by means of a simple formula, called the *gradient formula*, as shown in the following theorem.

Theorem 7.3 (Necessary conditions for differentiability). *Let $f : A \subseteq \mathbb{R}^2 \to \mathbb{R}$ (A open) and $(x_0, y_0) \in A$. If f is differentiable at (x_0, y_0), then it is continuous and has derivatives along every direction, given by the formula*

$$D_{\mathbf{v}} f(x_0, y_0) = \nabla f(x_0, y_0) \cdot \mathbf{v} = f'_x(x_0, y_0)\cos\alpha + f'_y(x_0, y_0)\sin\alpha. \tag{10.15}$$

Example 7.2. Let us compute the derivative of the function

$$f(x, y) = x^2 y$$

at the point $(1, 2)$, in the direction of the unit vector $\mathbf{v} = \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)$. Setting

$$g(t) = f\left(1 + \frac{t}{2}, 2 + \frac{\sqrt{3}t}{2}\right) = \left(1 + \frac{t}{2}\right)^2 \left(2 + \frac{\sqrt{3}t}{2}\right),$$

we have

$$g'(t) = \left(1 + \frac{t}{2}\right)\left(2 + \frac{\sqrt{3}}{2} + \frac{3\sqrt{3}t}{4}\right)$$

and then:

$$D_{\mathbf{v}} f(1, 2) = g'(0) = 2 + \frac{\sqrt{3}}{2}.$$

Since f is differentiable, we can also apply the gradient formula.
We have $\nabla f(x,y) = \left(2xy, x^2\right)$, so that

$$D_{\mathbf{v}} f(1,2) = \nabla f(1,2) \cdot \mathbf{v} = (4,1) \cdot \left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right) = 2 + \frac{\sqrt{3}}{2}.$$

From the formula (10.15), we can also deduce the important property:

Proposition 7.4. *The gradient of f at (x_0, y_0) indicates the direction of maximum growth*[7] *for f.*

10.7.1 The chain rule

Analogously to what happens in the one-dimensional case, expression (10.9) holds unchanged even if the variables x and y are differentiable functions of other variables. We recall that this property is called *invariance of form* of the first differential. We note immediately its usefulness through some important consequences, commonly used in economic theory.

• *A function of several variables, depending on time.* Let us suppose that the relation between the gross product Y and the availability of the production factors capital K and labour-power L in an economic system is known. We call it the production function, and we write

$$Y = F(K, L).$$

We assume now that the availabilities of both factors vary with time: $K = K(t)$, $L = L(t)$. Then the gross product Y is a function of time: $Y = F[K(t), L(t)]$. If F is differentiable with respect to K and L and these functions are differentiable with respect to t, then Y is also differentiable with respect to t and we can write

$$\begin{aligned} \mathrm{d}Y &= F'_K(K,L)\,\mathrm{d}K + F'_L(K,L)\,\mathrm{d}L = \\ &= F'_K(K,L)\,K'(t)\,\mathrm{d}t + F'_L(K,L)\,L'(t)\,\mathrm{d}t \end{aligned}$$

which is completely equivalent to

$$\frac{\mathrm{d}Y}{\mathrm{d}t} = F'_K(K,L)\,K'(t) + F'_L(K,L)\,L'(t).$$

In general, if $w(t) = F(x(t), y(t))$ we have the useful formula for the differentiation of composite functions (chain rule):

$$\boxed{\frac{\mathrm{d}w}{\mathrm{d}t} = \frac{\partial F}{\partial x}x'(t) + \frac{\partial F}{\partial y}y'(t).} \tag{10.16}$$

[7] The scalar product of two vectors attains its maximum when the two vectors are parallel and point in the same direction.

Another, and even more special, case we frequently meet concerns the relation between the gross product per worker $y = Y/L$ and the capital available per worker $k = K/L$, which are both functions of time. There are good reasons for assuming that y and k are connected by a relation of the type $y = f(k)$, called the *intensive production function*. The presence of technical progress makes such a relation between y and k vary in time:

$$y = f(k, t).$$

If we want to compute the derivative of y with respect to time, we must consider the fact that time has an influence on the product per worker both directly (because of the technical progress, through the second argument of the function) and indirectly (through the capital per worker k); therefore we have $y(t) = f[k(t), t]$. By using relation (10.16) with $x(t) = k(t)$ and $y(t) = t$, we get:

$$\frac{\mathrm{d}y}{\mathrm{d}t} = f'_k(k, t)\, k'(t) + f'_t(k, t).$$

10.8 Implicit functions

An equation such as

$$g(x, y) = 0, \tag{10.17}$$

satisfied at least at one point of the plane (x_0, y_0), usually identifies a curve through that point. A strategy for analysing such a curve is to try to establish whether, at least "piece-wise" (that is, locally), it may be expressed as the graph of a function like $y = f(x)$ or like $x = h(y)$. This way we can hope to use all the methods we illustrated in Chapter 6. If this is possible, the functions f and h are said to be *implicitly defined by the equation* (10.17) or simply *implicit functions*. Obviously, it is also important to guarantee that our *implicit* function is determined without any ambiguity, that is in a unique way.

Let us fix our attention on the possibility that (10.17) implicitly defines $y = f(x)$; the other case is perfectly analogous. As a first example, just to get started, we consider the equation

$$x - y = 0,$$

which is satisfied at $(0, 0)$ and which corresponds to the bisector of the first and third quadrants. Such a straight line is the graph of the (very simple) identity function $y = f(x) = x$. Now let us consider the equation

$$x^2 + y^2 - 1 = 0 \tag{10.18}$$

which represents the circumference with centre at the origin and radius 1. This curve is *not* the graph of a single function $y = f(x)$. If we set $x = 1/2$, for example, the value of y is not uniquely determined by equation (10.18), which reduces to

$$\left(\frac{1}{2}\right)^2 + y^2 - 1 = 0$$

and gives the two solutions $\pm\sqrt{3}/2$. Following the same steps of the previous example, we get two explicit formulae for f

$$f(x) = \begin{cases} \sqrt{1-x^2} & \text{(upper semi-circumference)} \\ -\sqrt{1-x^2} & \text{(lower semi-circumference).} \end{cases}$$

How can we choose one of the two? If we decide to choose the function through the point $(0, 1)$, we can exclude the lower semi-circumference and then consider the function

$$y = f(x) = \sqrt{1-x^2}, \tag{10.19}$$

as implicitly defined by equation (10.18).

However, we did not eliminate the ambiguity in the choice of the implicit function. In fact, still maintaining the condition $f(0) = 1$, nothing prevents us from constructing a function using some pieces of the upper semi-circumference and some pieces of the lower one, as in Figure 10. These functions are also implicitly defined by (10.18). Of course, with this technique we can construct an infinite number of implicit functions, and all of them are non-continuous functions. If we want to eliminate the ambiguity and select only one implicit function from our initial equation, we must limit ourselves to *continuous functions*.

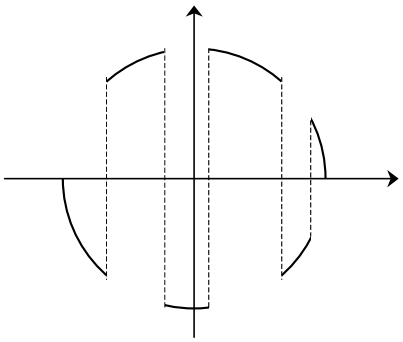


Figure 10.10. A non-continuous implicit function defined by equation (10.18)

In general, an equation $g(x, y) = 0$ cannot be solved by explicit formulae such as (10.19). For example, the equation

$$g(x, y) = x + \ln x - y - \ln y - 1 - \ln 2 = 0 \tag{10.20}$$

is satisfied at the point $(2, 1)$, but we are not able to get an explicit expression of y as a function of x or an explicit expression of x as a function of y. By means of a mathematical software like Maple, Matlab, Mathcad or similar ones, we can draw the locus corresponding to equation (10.20); the result is shown in Figure 11.

There are no problems in recognizing it as the graph of a continuous and differentiable function $y = f(x)$, implicitly defined by the equation $g(x, y) = 0$, but we are not able to write its analytic expression. What analytic information can we

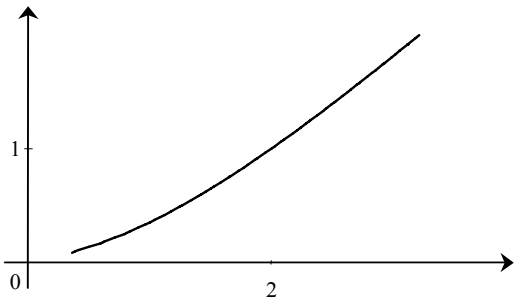


Figure 10.11. Implicit function defined by (10.20)

extract from (10.20)? Surely the most interesting information is the computation of the derivative of the implicit function $y = f(x)$ at the point $(2, 1)$, that is the point around which we are moving. In fact, the curve of equation (10.17) can be thought of as the zero-level curve of the function of two variables $z = g(x, y)$. Now, according to the remark on page 329, if

$$\nabla g(x_0, y_0) \neq \mathbf{0},$$

the tangent line to $g(x, y) = 0$ at (x_0, y_0) is the straight line of equation

$$g'_x(x_0, y_0)(x - x_0) + g'_y(x_0, y_0)(y - y_0) = 0.$$

If $g'_y(x_0, y_0) \neq 0$, we can write

$$y - y_0 = -\frac{g'_x(x_0, y_0)}{g'_y(x_0, y_0)}(x - x_0).$$

The slope of such a straight line is exactly the first derivative of the function $y = f(x)$ implicitly defined by $g(x, y) = 0$, computed at the point x_0. We have, therefore,

$$f'(x_0) = -\frac{g'_x(x_0, y_0)}{g'_y(x_0, y_0)}. \tag{10.21}$$

In the case of the implicit function defined by (10.20), we have $(x_0, y_0) = (2, 1)$ and:

$$g'_x(x, y) = 1 + \frac{1}{x}, \qquad g'_y(x, y) = -1 - \frac{1}{y},$$

$$g'_x(2, 1) = \frac{3}{2}, \qquad g'_y(2, 1) = -2 \neq 0.$$

Therefore, for the implicit function $y = f(x)$ we have

$$f'(2) = -\frac{g'_x(2, 1)}{g'_y(2, 1)} = \frac{3}{4}.$$

We can prove that the condition $g'_y(x_0, y_0) \neq 0$, with other hypotheses which are usually satisfied, guarantees the validity of the formula (10.21) not only at the point x_0, but also in a whole neighbourhood of the point itself. More precisely, the following theorem by Ulisse Dini (1845-1918) holds.

Theorem 8.1. *Let (x_0, y_0) be a point such that $g(x_0, y_0) = 0$. If*

(i) in a neighbourhood of (x_0, y_0) there exist the partial derivatives of g and they are continuous,

(ii) $g'_y(x_0, y_0) \neq 0$,

then the equation (10.17) *defines, in a neighbourhood of x_0, a unique implicit function $y = f(x)$, locally differentiable, with a continuous derivative given by the formula*

$$\boxed{\frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{g'_x(x, y)}{g'_y(x, y)}.} \tag{10.22}$$

In order to deduce (10.22), we can use the invariance of the first differential in the following way. If we imagine moving along the curve (10.17), the differential of g is equal to zero; in other words, if we imagine moving tangentially to the curve by an infinitesimal increment $(\mathrm{d}x, \mathrm{d}y)$, we have

$$g'_x(x, y)\,\mathrm{d}x + g'_y(x, y)\,\mathrm{d}y = 0. \tag{10.23}$$

Consequently, if $g'_y(x_0, y_0) \neq 0$, at least in a neighbourhood of (x_0, y_0):

$$\frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{g'_x(x, y)}{g'_y(x, y)}.$$

The relation (10.23), besides the undeniable advantage of being easy to memorize, does not favour either of the two variables x and y and reveals that, if we had the hypothesis $g'_x(x_0, y_0) \neq 0$ instead of $g'_y(x_0, y_0) \neq 0$, we could have obtained an implicit function of the type $x = h(y)$ with

$$\frac{\mathrm{d}x}{\mathrm{d}y} = -\frac{g'_y(x, y)}{g'_x(x, y)}.$$

When $\nabla g(x_0, y_0) = \mathbf{0}$, we need a more careful analysis, which we shall not deal with here.

Dini's theorem is often known, in the economic literature, as the *fundamental theorem of comparative statics*. We emphasize its relevance in such a context with an example.

• *How a supply curve is born.* Suppose that a firm sells x tons of goods produced on a wide market which expresses the price p, independent of x. Suppose that our firm works with increasing marginal costs: let the total cost of production be $C(x)$, with positive first and second derivatives, so that not only the total costs C, but also the marginal ones C' are increasing. The profit for the firm as a function of the quantity which is produced and sold is

$$\pi(x) = px - C(x),$$

with first derivative

$$\pi'(x) = p - C'(x)$$

and second derivative

$$\pi''(x) = -C''(x).$$

If the first derivative is null at a point x^*, that is [8]

$$p - C'(x^*) = 0 \tag{10.24}$$

the hypotheses concerning C'' imply that π'' is negative everywhere and therefore x^* is a global maximum point. The equation (10.24) binds the optimum amount offered to the market price, but, since we know C only from a qualitative point of view (through the sign of C' and C''), we certainly cannot hope to obtain an explicit x^* as a function of p. The function

$$x^* = f(p),$$

implicitly defined by the equation (10.24), is known as the *supply function* of the goods produced by the firm. We show that, under the mentioned hypotheses, the *supplied quantity increases with price.* It is sufficient to compute the sign of

$$\frac{\mathrm{d}x^*}{\mathrm{d}p} = -\frac{\mathrm{D}_p(p - C'(x^*))}{\mathrm{D}_{x^*}(p - C'(x^*))} = -\frac{1}{-C''(x^*)} = \frac{1}{C''(x^*)} > 0.$$

We note that we were able to reach this conclusion *even without knowing x^*, that is the value of the implicit function.*

10.9 Second order Taylor's formula

10.9.1 Second derivatives and Hessian matrix

We saw that computing the differential of a two-variable function leads us to determine the plane which is tangent to its graph. The information we expect to get from this knowledge is the same we get from the tangent line to the graph in one dimension. From the tangent plane we can understand for example if, slightly moving on the surface *in a given direction,* we are going up or down. By analogy with the one-dimensional case, we must look for extremum points among all points having a horizontal tangent plane, called *stationary* points. Thus we obtain a generalization of Fermat's necessary condition. The next step is to analyze the nature of a stationary point, in order to decide whether it is an extremum point or not and of which kind (maximum or minimum). In one dimension we saw that this kind of information is contained in Taylor's formula. In particular, Taylor's formula shows

[8] This condition should be considered with attention, because it translates the marginalistic optimum principle: *price = marginal cost.*

how the first order approximation for $\Delta f = f(a+h) - f(a)$, given by $f'(a)\,h$, can be significantly improved by

$$f'(a)\,h + \frac{1}{2} f''(a)\,h^2.$$

If $f''(a)$ is not null, the quantity $f''(a)\,h^2$, called the second differential of f at a, provides us with good information about the difference between the graph of f and the tangent line, near to the point a. Setting $h = x - a$, the parabola of equation

$$y = f(a) + f'(a)(x-a) + \frac{1}{2} f''(a)(x-a)^2, \tag{10.25}$$

is not only tangent to the graph of f at a, but it is also such that its slope at that point varies exactly as the slope of the graph.

These considerations can be extended to the case of functions of two (or more) variables; first of all, we must understand what we need to use instead of $f''(a)$. A function f of two variables has two first derivatives f'_x and f'_y. If we differentiate a second time, we can operate on each of the first derivatives with respect to each of the variables. Differentiating f'_x we obtain two second derivatives

$$f''_{xx} \quad \text{and} \quad f''_{xy},$$

the first one by differentiating again with respect to x, the second one by differentiating with respect to y. They can be also denoted by the symbols

$$\frac{\partial^2 f}{\partial x^2} \quad \text{and} \quad \frac{\partial^2 f}{\partial y \partial x}$$

respectively. By analogy, f'_y generates the two second derivatives

$$f''_{yx} = \frac{\partial^2 f}{\partial x \partial y} \quad \text{and} \quad f''_{yy} = \frac{\partial^2 f}{\partial y^2}.$$

Example 9.1. The gradient of the function

$$f(x,y) = x + 2y^2 + xy$$

is

$$\nabla f(x,y) = \left[\begin{array}{cc} 1+y & 4y+x \end{array}\right].$$

The four second derivatives are

$$\begin{aligned} f''_{xx} &= \frac{\partial}{\partial x}(1+y) = 0, \qquad f''_{xy} = \frac{\partial}{\partial y}(1+y) = 1, \\ f''_{yx} &= \frac{\partial}{\partial x}(4y+x) = 1, \quad f''_{yy} = \frac{\partial}{\partial y}(4y+x) = 4. \end{aligned}$$

We arrange the four partial derivatives in the matrix

$$\left[\begin{array}{cc} f''_{xx}(x,y) & f''_{xy}(x,y) \\ f''_{yx}(x,y) & f''_{yy}(x,y) \end{array}\right],$$

which is called the **Hessian matrix**[9] and is denoted by one of the symbols $f''(x,y)$ or $\nabla^2 f(x,y)$ or also $\mathbf{H}_f(x,y)$. The second derivatives with respect to the same variable, arranged on the main diagonal of the Hessian matrix, are said to be *pure*, the others are *mixed.*

Example 9.2. The Hessian matrix of the function we have seen in the previous example is

$$\nabla^2 f(x,y) = \begin{bmatrix} 0 & 1 \\ 1 & 4 \end{bmatrix} \tag{10.26}$$

The Hessian matrix of the function $g(x_1,x_2) = e^{2x_1+x_2}$ is

$$\nabla^2 g(x_1,x_2) = \begin{bmatrix} 4e^{2x_1+x_2} & 2e^{2x_1+x_2} \\ 2e^{2x_1+x_2} & e^{2x_1+x_2} \end{bmatrix}, \tag{10.27}$$

and therefore the Hessian matrix of this function changes from point to point.

The two mixed derivatives in (10.26) and (10.27) are equal. This is not a coincidence: it happens systematically, whenever such derivatives are continuous. Actually, the following theorem holds, so that the *symmetry* of the Hessian matrix is usually guaranteed.

Theorem 9.1 (Schwarz's theorem). *If the second derivatives of f are continuous, then their value does not depend on the order of differentiation*:

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}.$$

Since we are usually dealing with elementary functions and their composition and since their derivatives of every order are functions of the same type, we know *a priori* that they are continuous. Schwarz's theorem then allows us to save time in computations and the Hessian matrix turns out to be *symmetric.*

10.9.2 Second differential and second order Taylor's formula

The Hessian matrix $\nabla^2 f$ is, in some sense, the two-dimensional analogue of the second derivative f'', since the quadratic form associated with such a matrix takes a precise role in the second order Taylor's formula, as we shall soon see. In the meantime, we say that a function f is *twice differentiable at a point* (x_0,y_0) *if its first partial derivatives are differentiable at that point.* From theorem 7.2 (applied to the first derivatives) we know that this is true, for example, if all second derivatives are continuous at (x_0,y_0). In this case, the quadratic form

$$(h,k) \longmapsto f''_{xx}(x_0,y_0)\,h^2 + 2f''_{x,y}(x_0,y_0)\,hk + f''_{yy}(x_0,y_0)\,k^2 \tag{10.28}$$

takes the name of **second differential** of f at (x_0,y_0) and is denoted by the symbol $\mathrm{d}^2 f(x_0,y_0)$. We note that the form (10.28) may be written as follows using matrices:

$$(h,k) \to \begin{bmatrix} h & k \end{bmatrix} \nabla^2 f(x_0,y_0) \begin{bmatrix} h \\ k \end{bmatrix}.$$

[9] From the name of the German mathematician Ludwig Otto Hesse (1811-1874).

Now we have all the ingredients for *Taylor's formula.*

Theorem 9.2. *If $f : A \subseteq \mathbb{R}^2 \to \mathbb{R}$ has continuous second derivatives at the point (x_0, y_0), then the following second order Taylor's formula holds*:

$$\begin{aligned} f(x_0+h, y_0+k) &= f(x_0,y_0) + f'_x(x_0,y_0)\,h + f'_y(x_0,y_0)\,k + \\ &+\frac{1}{2}\left\{f''_{xx}(x_0,y_0)\,h^2 + 2f''_{x,y}(x_0,y_0)\,hk + f''_{yy}(x_0,y_0)\,k^2\right\} + \\ &+o\left(h^2+k^2\right) \qquad \text{for } (h,k)\to(0,0)\,. \end{aligned}$$

We explicitly note that the mere existence of the second derivatives does not allow us to call the quadratic function (10.28) second differential. But if the second derivatives are also continuous at (x_0, y_0), then f is twice differentiable and the name is correct.

10.10 Functions of n variables

All the notions introduced till now may be extended without problems to the case of functions of n variables, with $n \geq 3$. We present the definition of partial derivative and differential.

Definition 10.1. *Let $f : A \subseteq \mathbb{R}^n \to \mathbb{R}$, $\mathbf{a} \in A$. The **partial derivative** of f at the point $\mathbf{a}$, with respect to the variable x_k, is defined as the derivative of*

$$f(\mathbf{x}) = f(x_1, \ldots, x_k, \ldots, x_n)\,,$$

computed as if x_k were the only variable and the others were constant.

In other words, letting h be the increment of the variable x_k, we put

$$f'_{x_k}(\mathbf{a}) = \lim_{h\to 0} \frac{f(a_1,\ldots,a_k+h,\ldots,a_n) - f(a_1,\ldots,a_k,\ldots,a_n)}{h}$$

if the limit exists and is finite. It may be also denoted by the symbols

$$\frac{\partial f}{\partial x_k}(\mathbf{a})\,, \qquad \mathrm{D}_{x_k} f(\mathbf{a})\,.$$

Once the partial derivatives at a point are defined, we move on to *partial derivative functions.* If $f'_{x_k}(\mathbf{x})$ exists for every $\mathbf{x} \in A$, the function $f'_{x_k} : A \subseteq \mathbb{R}^n \to \mathbb{R}$, which associates the value $f'_{x_k}(\mathbf{x})$ with every $\mathbf{x} \in A$, are defined. Partial derivatives are computed, in the most common cases, by using the rules we have previously seen.

Example 10.1. Let $h(x,y,z) = (x+2y)\,e^{x+z^2}$; we have

$$\begin{aligned} h'_x(x,y,z) &= e^{x+z^2} + (x+2y)\,e^{x+z^2}, \\ h'_y(x,y,z) &= 2e^{x+z^2}, \\ h'_z(x,y,z) &= 2z\,(x+2y)\,e^{x+z^2}. \end{aligned}$$

The vector of all partial derivatives

$$\left[\begin{array}{cccc} f'_{x_1}(\mathbf{a}) & f'_{x_2}(\mathbf{a}) & \cdots & f'_{x_n}(\mathbf{a}) \end{array}\right]$$

is called the *gradient* of f at $\mathbf{a}$ and is denoted by the symbols $f'(\mathbf{a})$ or $\nabla f(\mathbf{a})$.

The definition of differentiability takes the following form: *we say that f is differentiable at the point $\mathbf{a}$ if there exists a (row) vector $\mathbf{m} \in \mathbb{R}^n$, such that, for every vector $\mathbf{h} \in \mathbb{R}^n$ for which $\mathbf{a}+\mathbf{h} \in A$, the increment of f can be written as*

$$f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) = \mathbf{mh} + o(|\mathbf{h}|) \qquad \text{for } \mathbf{h} \to \mathbf{0}. \tag{10.29}$$

The *linear* function of $\mathbf{h}$

$$\mathbf{h} \longmapsto \mathbf{mh}$$

takes the name of *(first) differential* of f at the point $\mathbf{a}$ and is denoted by the symbol $\mathrm{d}f(\mathbf{a})$.

Theorem 10.1. *If f is differentiable at $\mathbf{a}$ then:*

(i) f is continuous at $\mathbf{a}$,

(ii) all partial derivatives of f exist at $\mathbf{a}$ and, moreover, the vector $\mathbf{m}$ which appears in (10.29) *coincides with the gradient of f at* **a:**

$$\mathbf{m} = \left[\begin{array}{cccc} f'_{x_1}(\mathbf{a}) & f'_{x_2}(\mathbf{a}) & \cdots & f'_{x_n}(\mathbf{a}) \end{array}\right].$$

Then the product $\mathbf{mh}$ may be written as

$$\sum_{s=1}^{n} \frac{\partial f(\mathbf{a})}{\partial x_s} h_s,$$

and, if we introduce the differentials $\mathrm{d}x_s$ ("infinitesimal" increments) of the independent variables, we can write the differential as

$$\boxed{\mathrm{d}f(\mathbf{a}) = \sum_{s=1}^{n} \frac{\partial f(\mathbf{a})}{\partial x_s} \mathrm{d}x_s.}$$

Theorem 10.2 (Sufficient condition for differentiability). *If the partial derivatives of f exist in a neighbourhood of $\mathbf{a}$ and are continuous at the point $\mathbf{a}$ then f is differentiable at $\mathbf{a}$.*

Taylor's formula

For functions of n variables with continuous second derivatives, Taylor's formula takes the following form:

$$f(\mathbf{a}+\mathbf{h}) = f(\mathbf{a}) + \nabla f(\mathbf{a})\,\mathbf{h} + \frac{1}{2}\mathbf{h}^{\mathrm{T}} \nabla^2 f(\mathbf{a})\,\mathbf{h} + o\left(|\mathbf{h}|^2\right) \quad \text{for } \mathbf{h} \to \mathbf{0}$$

where $\nabla^2 f(\mathbf{a}) = \left[f_{x_j x_k}(\mathbf{a})\right]$ is the Hessian matrix. The quadratic form

$$\mathbf{h} \longmapsto \mathbf{h}^{\mathrm{T}} \nabla^2 f(\mathbf{a})\,\mathbf{h}$$

is the *second differential of f at the point* $\mathbf{a}$ and is denoted by the symbol $\mathrm{d}^2 f(\mathbf{a})$.

10.11 Optimization. Unconstrained extrema

10.11.1 Unconstrained and constrained extrema

The analysis of Economics and Business problems frequently requires the introduction of functions $y = f(\mathbf{x}) = f(x_1, x_2, \ldots, x_n)$, which describe the result y of a decision concerning the choice of $\mathbf{x}$ in a set $A \subseteq \mathbb{R}^n$. In many cases we are interested in the vectors $\mathbf{x}^*$ maximizing or minimizing y. Two examples help us understand the essence of the problem, which may occur in two different forms.

• If $x_1, x_2, \ldots, x_n$ are the amounts of n goods a firm can produce, and if the productivity of the firm is quite high, the usual obstacles to a significant growth in the production of the n goods are typically market obstacles: in order to sell all the product, prices should be very low. Then it might be a good idea to appropriately contain the volumes of production in order to optimize the result (for instance, maximize the profit obtained by the firm). In this case, given a certain vector $\mathbf{x}^*$ of the produced amounts, $\mathbf{x}^*$ can be changed in any possible way: we can increase the production of all goods, decrease the production of all goods, increase the volume of some of them and decrease the volume of the others, without meeting any physical, economic or financial obstacle. Changing $\mathbf{x}^*$ has the only consequence of a (desirable) rise or a (deplorable) reduction of the result.

• A firm works with scarce resources and the vector which presently describes the production of the n goods requires the complete use of at least some of these resources (sometimes called *bottlenecks*). In this case the firm cannot vary $\mathbf{x}^*$ in any desired direction: probably a generalized increase of the production volumes would be impossible, because of the presence of bottlenecks. An increase in the volume of some of the goods might be possible under the condition that we reduce the volume of the others. We are still interested in maximizing the profit y originated by our decisions about production, but we must now take into account the shortage of some resources.

The first case is typical of *unconstrained optimization* problems. We are looking for the point $\mathbf{x}^*$ which maximizes (or minimizes) a function of n variables, supposing that we can freely move in a set A, which is an open set of $\mathbb{R}^n$. The points $\mathbf{x}$ which are candidates to be the optimum point are all interior points of A, and therefore we can move around them in all directions.

The second case is typical of *constrained optimization* problems (constrained by the presence of scarce resources). Not all directions of movement from the optimum points $\mathbf{x}$ are possible; the only possible ones are those compatible with the available amount of resources.

In this section we are interested in unconstrained optimization problems, restricting ourselves to the case of two-variable functions, with continuous second partial derivatives. In this case, the search for unconstrained (local) extrema may be accomplished in two steps.

— First, we determine the points where the gradient is equal to zero (first order condition). Such points are called *stationary;* from a geometric point of view, a null gradient indicates that the tangent plane at the corresponding point on the graph of the function is parallel to the x, y-plane.

— We then investigate the sign of the second differential at the stationary points (second order conditions).

10.11.2 First and second order conditions

Fermat's theorem

The first of the two steps we listed at the end of the previous section leads to a generalization of Fermat's theorem, the second one leads to the study of a quadratic form.

Theorem 11.1. *Let $f : A \subseteq \mathbb{R}^2 \to \mathbb{R}$ have partial derivatives at a point (x_0, y_0) interior for A. If (x_0, y_0) is a local extremum point for f then $\nabla f(x_0, y_0) = \mathbf{0}$.*

Proof. Let us fix y_0 and let x vary. Thus we obtain the function $F(x) = f(x, y_0)$, depending only on the variable x. Such a function has a local extremum point at x_0 and then, according to the one-dimensional Fermat's theorem, the derivative of F at x_0 must be zero. But, by definition, we have $f'_x(x_0, y_0) = F'(x_0)$ and therefore $f'_x(x_0, y_0) = 0$. Similarly, we can prove that $f'_y(x_0, y_0) = 0$. □

We have already said that the points where ∇f is null are called *stationary points.*

Example 11.1. Let

$$f(x, y) = x^2 - 2x + y^4 + y^2.$$

The gradient of f is

$$\nabla f(x, y) = \begin{bmatrix} 2x - 2 & 4y^3 + 2y \end{bmatrix}$$

and is null only for $x = 1$ and $y = 0$, thus $(1, 0)$ is the only stationary point of f. At the moment we cannot say anything more about the nature of this point. We need second order conditions.

Sign of the second differential

In the case of the optimization of one-variable functions, the useful information about the second derivative was included in its sign. In fact the expression $1/2 f''(a) h^2$, which appears in the one-dimensional Taylor's formula, is positive or negative (if $h \neq 0$) according to the sign of the second derivative. In the case of two-variable functions, we need to consider the sign of the second differential.

Let $a = f''_{xx}(x_0, y_0)$, $b = f''_{xy}(x_0, y_0) = f''_{yx}(x_0, y_0)$ and $c = f''_{yy}(x_0, y_0)$. The Hessian matrix of f at the point (x_0, y_0) is therefore

$$\nabla^2 f(x_0, y_0) = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

and the second differential at (x_0, y_0) is the quadratic form

$$ah^2 + 2bhk + ck^2. \tag{10.30}$$

Thus, in order to study the sign of $\mathrm{d}^2 f(x_0, y_0)$, we can use the results we saw in section 4. What kind of information can we obtain from the sign of the second differential? Let us consider a simple example.

Example 11.2. Let $f(x, y) = x^2 + 2y^2 - x^3$. Its gradient is

$$\nabla f(x, y) = \begin{bmatrix} 2x - 3x^2 & 4y \end{bmatrix}$$

which is null, for instance[10], at $(0, 0)$, where the function is null as well. The Hessian matrix of f is

$$\nabla^2 f(x, y) = \begin{bmatrix} 2 - 6x & 0 \\ 0 & 4 \end{bmatrix}$$

which at $(0, 0)$ becomes

$$\begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}.$$

This matrix is positive definite and the second order term is therefore positive. We can deduce that[11]

$$f(x, y) = 0 + 0 + x^2 + 2y^2 + o\left(x^2 + y^2\right) \quad \text{for } (x, y) \to (0, 0).$$

Thus, near to the origin, f takes positive values, greater than the value at $(0, 0)$, which is consequently a *local minimum* point.
Let us examine the example. The function f is written as sum of

— a positive quadratic form $(= x^2 + 2y^2)$ and

— a term which becomes smaller and smaller when we get close to the origin $(= -x^3)$.

Then we can consider f, near to the origin, as a small deformation of the quadratic form. The previous calculation shows that, since at the origin the quadratic form has a *strict minimum,* if we slightly change its shape nearby things do not substantially change. In the worst case, the minimum will no longer be global, but only local. The effect is illustrated in Figure 12.

The reader is invited to carry out other experiments with a computer, for example with the functions

$$f_1(x, y) = x^2 + 2y^2 - x^2 y, \qquad f_2(x, y) = x^2 + 2y^2 - xy^4,$$

obtained by adding terms of higher degree to the same quadratic form. The graphs will show that the origin is still a (local) minimum point.

Things are working well with positive (and negative) definite quadratic forms. If we try the same experiment on the quadratic form $g_1(x, y) = x^2$, which is still non-negative and vanishes at all vectors $(0, y)$, which are therefore weak minimum points, the result is different. In Figure 13 we show the graph of g_1 and of its deformation $g_2(x, y) = x^2 - y^3$. The origin is no longer an extremum point, not even locally.

[10] It is null also at $(2/3, 0)$.

[11] Since our Taylor's formula is centred at the origin, we will call it a Maclaurin's formula and we will identify the increment vector with (x, y).

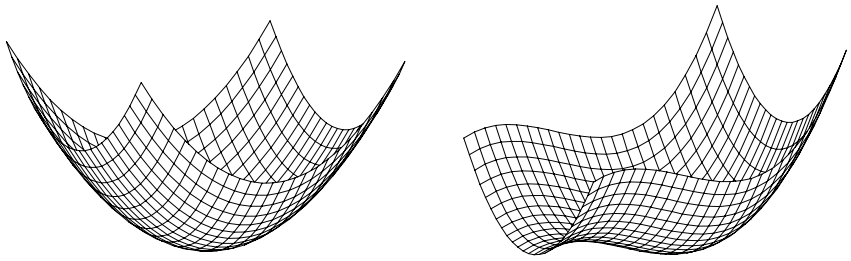


Figure 10.12.

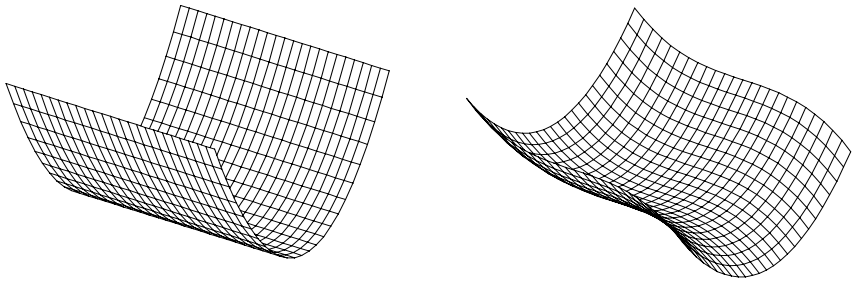


Figure 10.13.

These simple examples illustrate the main idea at the core of unconstrained optimization. If (x_0, y_0) is a stationary point, since $f'_x(x_0, y_0) = f'_y(x_0, y_0) = 0$, using Taylor's formula we can write, for $(h, k) \to (0, 0)$,

$$f(x_0 + h, y_0 + k) - f(x_0, y_0) =$$

$$= \frac{1}{2}\left\{f''_{xx}(x_0, y_0)\, h^2 + 2f''_{xy}(x_0, y_0)\, hk + f''_{yy}(x_0, y_0)\, k^2\right\} + o\left(h^2 + k^2\right).$$

We note that, near to the origin, the function

$$g(h, k) = f(x_0 + h, y_0 + k) - f(x_0, y_0)$$

turns out to be a small perturbation of the quadratic form

$$\frac{1}{2}\mathrm{d}^2 f(x_0, y_0) = \frac{1}{2}\left\{f''_{xx}(x_0, y_0)\, h^2 + 2f''_{xy}(x_0, y_0)\, hk + f''_{yy}(x_0, y_0)\, k^2\right\}.$$

If this form is positive definite, the same holds for g, which has consequently a strict minimum at the origin. This implies that f, in turn, has a strict minimum at (x_0, y_0). An analogous argument holds if $\mathrm{d}^2 f(x_0, y_0)$ is negative definite. On the other hand, if the second differential is an *indefinite* quadratic form (that is $\mathrm{d}^2 f$ changes its sign in every neighbourhood of $(0, 0)$), then the point (x_0, y_0) *cannot be an extremum.* The sign of $\mathrm{d}^2 f(x_0, y_0)$ may obviously be deduced from the Hessian matrix at (x_0, y_0).

Conclusions may be summarized in the following theorem, as a sufficient condition for the presence of a local extremum point. Since in this condition the second derivatives of f are involved, it is called a *second order* condition.

Theorem 11.2. *Let $f : A \subseteq \mathbb{R}^2 \to \mathbb{R}$ and (x_0, y_0) be a stationary point, interior for A, where f has continuous second partial derivatives. If the Hessian matrix $\nabla^2 f(x_0, y_0)$ is:*

(i) negative (positive) definite, then (x_0, y_0) is a strict local maximum (minimum) point;

(ii) indefinite, then (x_0, y_0) is not an extremum point.

The theorem allows us to reach a conclusion in all cases when the Hessian matrix is definite or indefinite. In the other cases (when the matrix is semi-definite) further investigations are needed, which we do not intend to explain here.

Let us go back to the example 11.1; the Hessian matrix of f is

$$\nabla^2 f(x, y) = \begin{bmatrix} 2 & 0 \\ 0 & 12y^2 + 2 \end{bmatrix}.$$

At the point $(1, 0)$ it becomes

$$\nabla^2 f(1, 0) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix},$$

which is positive definite: the point $(1, 0)$ is a strong local minimum point.

From theorem 11.2 we deduce the following rule:

If $\det\left(\nabla^2 f(x_0, y_0)\right) > 0$ *and*

$$\begin{cases} f''_{xx}(x_0, y_0) > 0, \text{ then } (x_0, y_0) \text{ is a strict local minimum point;} \\ f''_{xx}(x_0, y_0) < 0, \text{ then } (x_0, y_0) \text{ is a strict local maximum point.} \end{cases}$$

If $\det\left(\nabla^2 f(x_0, y_0)\right) < 0$ *, then (x_0, y_0) is a saddle point.*

If $\det\left(\nabla^2 f(x_0, y_0)\right) = 0$, the test of the Hessian matrix does not provide sufficient information and a deeper investigation is needed.

• *The method of least squares.* Many interesting problems in Economics and in Business Administration may be reduced to searching for a linear function

$$y = \beta x + \alpha$$

which fits as best as possible a sequence of observed data

$$(x_1, y_1), \quad (x_2, y_2), \ldots, \quad (x_n, y_n).$$

We are dealing with (many) values of y associated with values of x, which we shall suppose not to be all equal, in order not to deprive our problem of interest. For example, x could be the national income and y the amount of total consumptions. The pairs (x_i, y_i) are related to years $i = 1, 2, \ldots, n$. As an alternative, x could be the amount spent in advertising by a given firm in a given district and y could be the corresponding volume of sales.

In the study of stock markets we often try to estimate relationships of this kind, connecting y, which is the difference between the yield rate of one stock with respect to a riskless return (Treasury bills, for instance), with x, which is the difference between the yield rate of the market with respect to the riskless return. In Finance, the parameters appearing in the equation of the above straight line are indeed referred to as *beta* and *alpha coefficients.*

In both cases, it is natural to think that, at least approximately, y is made up of a term α which is independent of x (the consumption for surviving, in the first case; the guaranteed market share, in the second) and a term βx which is proportional to x, where β is known, in the first case, as the *marginal propensity to consumption* and, in the second, as the *marginal efficiency of advertising.* In the case of stock yield rates, β is used to classify stocks as *aggressive* (with $\beta > 1$, so that the variations of their yield rate are more relevant than the market average variations) and *defensive* (with $0 < \beta < 1$, so that the variations of their yield rates are smaller then the variations of the yield rates of the market).

Obviously, in none of the quoted cases we expect that the (affine) linear relation between y and x holds exactly, that is:

$$y_i = \beta x_i + \alpha \quad \text{for every } i.$$

We formulate the hypothesis that it holds, up to an error ε_i :

$$y_i = \beta x_i + \alpha + \varepsilon_i \quad \text{for every } i.$$

In other words, the "true" value y_i is equal to the theoretical value $\beta x_i + \alpha$, apart from a "noise" represented by the error ε_i. How can we determine the parameters α and β so that the terms "ε_i" are generally minimized? The first idea consists in minimizing the sum $\sum_{i=1}^{n} \varepsilon_i$, but we immediately realize that relevant errors with opposite sign might compensate and make the sum become small even for a straight line which does not really fit the data. The sum of the squares of such errors (obviously seen as functions of β and α) does not allow any compensation, and thus it is a better measure of the goodness of the approximation. Such a sum is given by

$$E(\beta, \alpha) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (\beta x_i + \alpha - y_i)^2,$$

and includes the squares of all "vertical distances" of the points (x_i, y_i) from the straight line. By the term "vertical distances" we mean the difference between the ordinate of the point $(x_i, \beta x_i + \alpha)$, belonging to the line, and that of the point (x_i, y_i).

Now the problem is to minimize the function $E(\beta, \alpha)$, which is defined on $\mathbb{R}^2$ with partial derivatives which are continuous at every point. In order to determine all its stationary points, we compute the partial derivatives of E and set them equal to zero. We have

$$\begin{cases} E'_\beta(\beta, \alpha) = \sum_{i=1}^{n} 2x_i(\beta x_i + \alpha - y_i) = 0 \\ E'_\alpha(\beta, \alpha) = \sum_{i=1}^{n} 2(\beta x_i + \alpha - y_i) = 0, \end{cases}$$

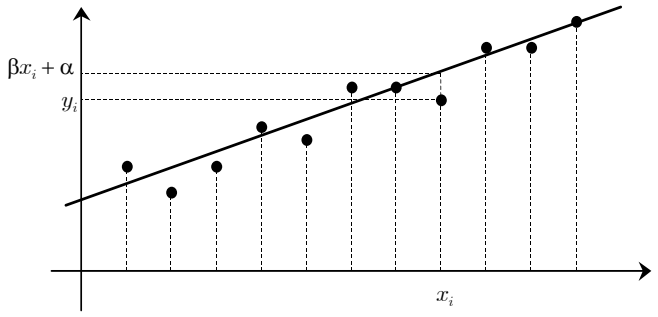


Figure 10.14. The straight line approximating a set of points

that is

$$\begin{cases} \beta \sum_{i=1}^{n} x_i^2 + \alpha \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ \beta \sum_{i=1}^{n} x_i + n\alpha = \sum_{i=1}^{n} y_i. \end{cases} \tag{10.31}$$

The system turns out to be linear in α and β. Let

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}; \quad \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}; \quad \overline{q} = \frac{\sum_{i=1}^{n} x_i^2}{n}; \quad \overline{p} = \frac{\sum_{i=1}^{n} x_i y_i}{n}.$$

The numbers $\overline{x}$, $\overline{y}$, $\overline{q}$ and $\overline{p}$ represent the arithmetical means of the abscissae, of the ordinates, of the squares of abscissae and of the products abscissa times ordinate, respectively. After dividing each term of the equations by n, the system (10.31) can be written as:

$$\begin{cases} \overline{q}\beta + \overline{x}\alpha = \overline{p} \\ \overline{x}\beta + \alpha = \overline{y}. \end{cases}$$

If not all of the x_is are equal[12], the determinant of the coefficient matrix of the system is

$$\begin{vmatrix} \overline{q} & \overline{x} \\ \overline{x} & 1 \end{vmatrix} = \overline{q} - \overline{x}^2 > 0,$$

therefore the system has the unique solution (β, α) given by:

$$\beta = \frac{\overline{p} - \overline{x} \cdot \overline{y}}{\overline{q} - \overline{x}^2}, \qquad \alpha = \frac{\overline{q} \cdot \overline{y} - \overline{p} \cdot \overline{x}}{\overline{q} - \overline{x}^2}.$$

As we can intuitively realize, the point (β, α), which is the only stationary point for the function E, is a global minimum point, since E is a positive quadratic function.

[12] Indeed, the positive number $\sum_{i=1}^{n} (x_i - \overline{x})^2$ may be easily re-written as

$$\sum_{i=1}^{n} x_i^2 - 2\overline{x} \sum_{i=1}^{n} x_i + n\overline{x}^2 = n\overline{q} - n\overline{x}^2 = n\left(\overline{q} - \overline{x}^2\right),$$

therefore we must have $\overline{q} > \overline{x}^2$. The quantity $\overline{q} - \overline{x}^2$ is called the *variance* of the data x_i and is widely used in Statistics.

The straight line of equation $y = \beta x + \alpha$ is the line we were looking for. The numerator in the expression of the slope β is known in Statistics as the *covariance* between x and y, while the denominator $\overline{q} - \overline{x}^2$ is called the *variance* of x. We have obtained the well-known formula which defines, in the financial theory, the beta coefficient of a stock:

$$\beta = \frac{\text{covariance between } x \text{ and } y}{\text{variance of } x}.$$

• *Functions of n variables.* Theorem 11.2 holds unchanged also for functions of any number of variables. For example, let us optimize the function

$$f(\mathbf{x}) = 3x_1^2 + 2x_1x_2 + x_2^2 + x_1x_3 + x_3^2 - x_1x_2^3.$$

We compute its gradient

$$\nabla f(\mathbf{x}) = \begin{bmatrix} 6x_1 + 2x_2 + x_3 - x_2^3 & 2x_1 + 2x_2 - 3x_1x_2^2 & x_1 + 2x_3 \end{bmatrix}$$

and we set it equal to zero:

$$\begin{cases} 6x_1 + 2x_2 + x_3 - x_2^3 = 0 \\ 2x_1 + 2x_2 - 3x_1x_2^2 = 0 \\ x_1 + 2x_3 = 0. \end{cases}$$

A solution for the system is $x_1 = x_2 = x_3 = 0$ and, consequently, the null vector $\mathbf{0}$ is a stationary point. In order to decide about its nature, we compute the Hessian matrix

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} 6 & 2 - 3x_2^2 & 1 \\ 2 - 3x_2^2 & 2 - 6x_1x_2 & 0 \\ 1 & 0 & 2 \end{bmatrix},$$

which at $\mathbf{0}$ becomes

$$\nabla^2 f(\mathbf{0}) = \begin{bmatrix} 6 & 2 & 1 \\ 2 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix}.$$

Its NW principal minors are

$$6, \quad \det \begin{bmatrix} 6 & 2 \\ 2 & 2 \end{bmatrix} = 8, \quad \det \begin{bmatrix} 6 & 2 & 1 \\ 2 & 2 & 0 \\ 1 & 0 & 2 \end{bmatrix} = 14.$$

Since the chain of signs is +,+,+, the origin is a strong local minimum.

10.12 Constrained extrema

10.12.1 Explicit constraint

We introduce constrained optimization problems with an example, which may be solved by means of elementary methods.

• *The producer problem.* We present a very elementary model of micro-economics. A firm invests amounts of capital and labour-power respectively equal to K and L, and obtains a product Y. The dependence of Y on K, L is described by a Cobb-Douglas production function

$$Y = f(K, L) = aK^{\alpha}L^{1-\alpha}, \quad \text{with } a > 0,\ 0 < \alpha < 1.$$

The problem of maximizing Y without any constraint is meaningless, both in the common sense and in the economic sense. Since f is indefinitely increasing as the invested quantities of each factor increase (provided that both quantities are positive), it attains no maximum, and this makes sense: if we can indefinitely expand the investment of each production factor, the product obtained increases indefinitely. From an economic point of view, an interesting problem is to maximize Y, while satisfying some constraints on the production factors. It is reasonable to think that the firm has a given budget $b > 0$, to be used in order to purchase capital services and labour-power. One unit of capital costs p_K, one unit of labour-power p_L, so that the cost for purchasing the amounts (K, L) turns out to be $Kp_K + Lp_L$. Let us suppose that the cost cannot exceed the budget. Here is the constraint which makes the question economically interesting: what pair (K^*, L^*) maximizes Y, under the budget condition $Kp_K + Lp_L \leq b$? It is reasonable to assume that it is not profitable to under-use the budget, thus the optimum pair (K^*, L^*) which maximizes Y will satisfy the condition of exhaustion of the budget; in this case we say the constraint on the budget is "saturated": $Kp_K + Lp_L = b$. Therefore we are interested in solving the problem which may be codified as follows:

$$\begin{cases} \max_{K,L} f(K, L) = aK^{\alpha}L^{1-\alpha} \\ sub \\ Kp_K + Lp_L = b. \end{cases}$$

How can we solve the problem? Since the constraint equation may be explicitly solved with respect to K or L, we are led back to an unconstrained problem for a one-variable function. Deducing from the constraint

$$K = \frac{b - Lp_L}{p_K},$$

we have to maximize the function of the only variable L

$$h(L) = a\left(\frac{b - Lp_L}{p_K}\right)^{\alpha} L^{1-\alpha} = \frac{a}{p_K^{\alpha}}(b - Lp_L)^{\alpha} L^{1-\alpha}$$

in the interval $0 < L < b/p_L$. Setting h' equal to zero, we obtain the equation

$$-\alpha p_L (b - Lp_L)^{\alpha-1} L^{1-\alpha} + (1 - \alpha)(b - Lp_L)^{\alpha} L^{-\alpha} = 0.$$

Simplifying, we have

$$-\alpha p_L L + (1 - \alpha)(b - Lp_L) = 0.$$

This condition gives us the solution for L and consequently for K:

$$L^* = \frac{b(1-\alpha)}{p_L} \qquad K^* = \frac{b\alpha}{p_K}.$$

The pair of values we have found indeed corresponds to a maximum point, as we can easily see from the change in the sign of h' at L^*.

The method we have followed only works when constraints can be made explicit with respect to one of the variables, and anyone can understand that this rarely happens, especially if the variables are more than two.

Indeed the problem we have seen can be generalized. Let us consider an *object function* $f : \mathbb{R}^n \to \mathbb{R}$

$$f(x_1, x_2, \ldots, x_n) = f(\mathbf{x}),$$

a vector of m functions $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), ..., g_m(\mathbf{x})]^{\mathrm{T}}$, and a vector $\mathbf{b}$ with m components; we look for a vector $\mathbf{x}^*$ with n components which maximizes f while satisfying the *constraint* conditions

$$\mathbf{g}(\mathbf{x}^*) = \mathbf{b}. \tag{10.32}$$

Obviously, the formulation of the problem is reasonable only if m, which is the number of constraint equations, is *lower* than n, the number of variables. If the equality $m = n$ held, the condition (10.32) could be satisfied, in general, at most by a finite number of points.

This problem has an interesting economic meaning. A firm produces n goods, whose amounts are $\mathbf{x} = [x_1, x_2, \ldots, x_n]^{\mathrm{T}}$, and the function f indicates the contribution margin[13] $f(\mathbf{x})$, determined by the *production mix* $\mathbf{x}$. The production process of the firm fully uses m resources, available in the quantities $\mathbf{b} = [b_1, b_2, \ldots, b_m]^{\mathrm{T}}$. The problem which is indicated by

$$\begin{cases} \max_{\mathbf{x}} f(\mathbf{x}) \\ sub \\ \mathbf{g}(\mathbf{x}) = \mathbf{b}. \end{cases}$$

expresses the micro-economic problem of the producer who aims at optimizing (through the maximization of $y = f(\mathbf{x})$) the use of resources at his disposal (collected in the vector $\mathbf{b}$), which are absorbed as described by the function $\mathbf{g}(\mathbf{x})$.

10.12.2 Lagrange multipliers

Referring to the case of the *production mix*, we shall now study the case of two kinds of goods ($n = 2$, that is $\mathbf{x} = [x, y]^{\mathrm{T}}$) and one resource at our disposal ($m = 1$, $\mathbf{b} = b$, $g : \mathbb{R}^2 \to \mathbb{R}$). The problem is then

$$\mathcal{P}_1 := \begin{cases} \max_{x,y} f(x, y) \\ sub \\ g(x, y) = b, \end{cases}$$

[13] This is the difference between sales revenue and variable costs: it is the quantity a firm is interested in maximizing, when the structure of fixed costs is given.

where f and g are meant to be differentiable in an open set A. We shall say that a constrained maximum point (x^*, y^*) for f *is a solution of the problem* $\mathcal{P}_1$. As in the case of unconstrained optimization, first of all we look for first order necessary conditions for the presence of an optimum point. Suppose therefore that the point (x^*, y^*) is a solution of the problem $\mathcal{P}_1$. The following theorem holds.

Theorem 12.1. *Let f and g be differentiable and $\nabla g(x^*, y^*) \neq \mathbf{0}$. If the point (x^*, y^*) is a solution of the problem $\mathcal{P}_1$, then there exists a number λ^* such that*

$$\nabla f(x^*, y^*) = \lambda^* \nabla g(x^*, y^*). \tag{10.33}$$

Condition (10.33) reveals that *at the point (x^*, y^*) the gradients of f and g are parallel.* The geometric meaning of such a condition is interesting.

Let us consider the constraint curve $b - g(x, y) = 0$ and the level curve $f(x, y) = f(x^*, y^*)$. As we saw in section 7, the vector $\nabla g(x^*, y^*)$ is orthogonal to the constraint, while $\nabla f(x^*, y^*)$ is orthogonal to the level curve of f. Then, from (10.33), we deduce that, if the gradients are not null at (x^*, y^*), the two curves *are tangent* (see Figure 15).

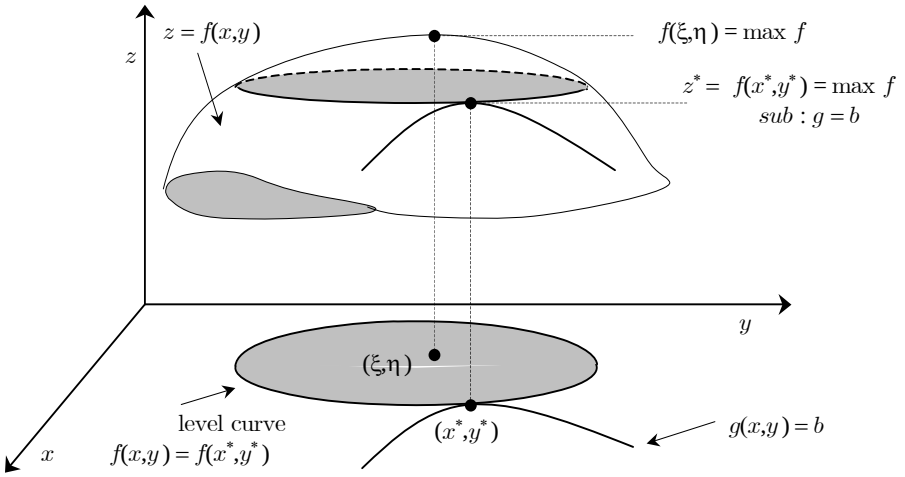


Figure 10.15.

If we introduce the *Lagrangean function*

$$L(\lambda, x, y) = f(x, y) + \lambda [b - g(x, y)],$$

theorem 12.1 is equivalent to stating that: *the point (λ^*, x^*, y^*) is a stationary point for L, that is*

$$\nabla L(\lambda^*, x^*, y^*) = \mathbf{0}.$$

The function L is usually simply called the *Lagrangean* of the problem; λ is called the *Lagrange multiplier*. The condition for stationary points $\nabla L(\lambda, x, y) = \mathbf{0}$ is

equivalent to the following system:

$$\begin{cases} L'_\lambda(\lambda, x, y) = b - g(x, y) = 0 \\ L'_x(\lambda, x, y) = f'_x(x, y) - \lambda g'_x(x, y) = 0 \\ L'_y(\lambda, x, y) = f'_y(x, y) - \lambda g'_y(x, y) = 0. \end{cases} \tag{10.34}$$

It is a system of three equations in the three unknowns λ, x, y. The first equation is nothing but the constraint condition, returned to us by the requirement for a stationary Lagrangean with respect to the multiplier. The last two equations may be written as $\nabla f(x, y) = \lambda \nabla g(x, y)$, which is the condition (10.33) evaluated at (λ^*, x^*, y^*).

The necessary condition for optimum points expressed in theorem 12.1 does not make any distinction between maximum or minimum points. Indeed, a minimum problem for the function f is equivalent to a maximum problem for $-f$, provided that the constraint remains the same. The first order necessary condition is identical. Now suppose that we have solved the system (10.34), finding the values λ^*, x^*, y^* for the three unknowns. Two questions naturally arise so far.

(a) How can we check that the point (x^*, y^*), found through the system (10.34), is really a maximum or a minimum point?

(b) What is the meaning of λ^*?

We immediately offer a partial answer to the first question; later on, we shall discuss the meaning of the multiplier.

Theorem 12.2. *Let $\nabla L(\lambda^*, x^*, y^*) = \mathbf{0}$. If L has continuous second derivatives and the determinant of its Hessian matrix is positive (negative) at (λ^*, x^*, y^*), then the point (x^*, y^*) is a constrained local maximum (minimum) point.*

Example 12.1. Let

$$f(x, y) = x^2 + y^2,$$

to be minimized under the constraint $x + y = 10$.

In this problem $b = 10$, $g(x, y) = x + y$. The constraint may be written explicitly: the reader is invited to solve the exercise using this fact. Applying the method of multipliers, we write the Lagrangean of the problem:

$$L(\lambda, x, y) = x^2 + y^2 + \lambda(10 - x - y)$$

so that the system (10.34) is:

$$\begin{cases} L'_\lambda = 10 - x - y = 0 \\ L'_x = 2x - \lambda = 0 \\ L'_y = 2y - \lambda = 0. \end{cases}$$

From the last two equations we deduce $x^* = y^*$; for the constraint condition, their value can only be 5. We have also $\lambda^* = 10$.

The Hessian matrix of the Lagrangean is

$$L'' = \begin{bmatrix} 0 & -1 & -1 \\ -1 & 2 & 0 \\ -1 & 0 & 2 \end{bmatrix},$$

and its determinant is $-4 < 0$, therefore $f(5,5)$ is a local (actually, also global) constrained minimum.

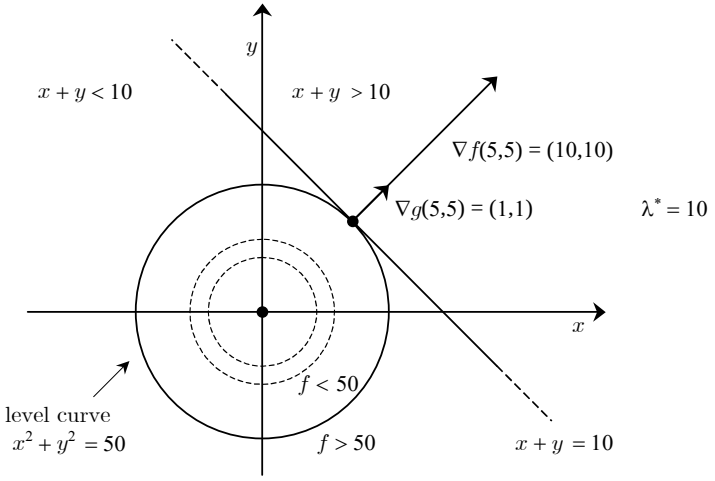


Figure 10.16.

Figure 16 geometrically illustrates what happens. The object function is nothing but the square of the distance of the point (x, y) from the origin. The problem consists in determining the point at which the straight line $x + y = 10$ (the constraint) is tangent to the level curve of f, that is the circumference $x^2 + y^2 = 50$.

The positive sign of the multiplier agrees with the fact that the global minimum of f is at the origin, which belongs to the region where $x + y < 10$: the gradients of f and g at the point $(5, 5)$ point in the same direction.

Example 12.2. Theorem 12.2 provides information when the optimum point (x^*, y^*) is a regular point for the constraint, that is when $\nabla g(x^*, y^*)$ is not null. However, let us try and solve the problem

$$\begin{cases} \min\left(x^2 + y^2\right) \\ sub \\ g(x_1, x_2) = x^2 - (y-1)^3 = 0. \end{cases}$$

From a geometric point of view, we have to find the point which satisfies the constraint and has the shortest distance from the origin. Figure 17 shows that the point we are looking for exists and is $(0, 1)$. Nevertheless $\nabla f(0,1) = \begin{bmatrix} 0 & 2 \end{bmatrix}$ while $\nabla g(0,1) = \begin{bmatrix} 0 & 0 \end{bmatrix}$, so that the system (10.34) turns out to be impossible. There are no stationary points for the Lagrangean.

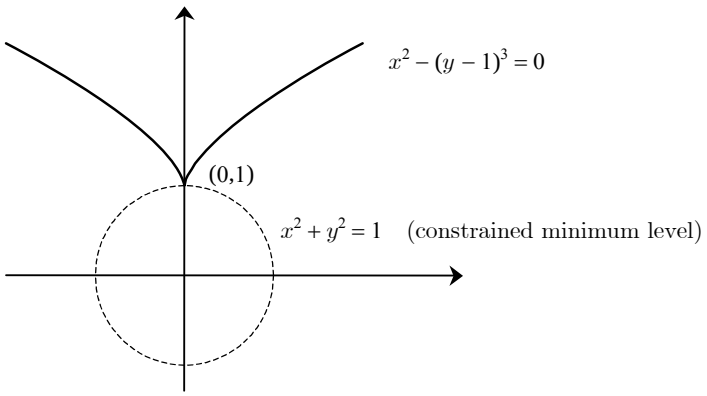


Figure 10.17.

• *Maximum production under a budget constraint.* With the method of multipliers we solve the problem

$$\begin{cases} \max_{K,L} aK^{\alpha}L^{(1-\alpha)} \\ sub \\ Kp_K + Lp_L = b \end{cases}$$

we met on page 350. The Lagrangean is

$$L(\lambda, K, L) = aK^{\alpha}L^{(1-\alpha)} + \lambda(b - Kp_K - Lp_L),$$

with gradient

$$L' = \left[\begin{array}{ccc} b - Kp_K - Lp_L & a\alpha K^{\alpha-1}L^{1-\alpha} - \lambda p_K & a(1-\alpha)K^{\alpha}L^{-\alpha} - \lambda p_L \end{array} \right].$$

The first order conditions lead to the system

$$\begin{cases} b - Kp_K - Lp_L = 0 \\ a\alpha K^{\alpha-1}L^{1-\alpha} - \lambda p_K = 0 \\ a(1-\alpha)K^{\alpha}L^{-\alpha} - \lambda p_L = 0, \end{cases}$$

whence, eliminating the Lagrange multiplier between the last two equations, we get

$$\frac{a\alpha K^{\alpha-1}L^{1-\alpha}}{p_K} = \frac{a(1-\alpha)K^{\alpha}L^{-\alpha}}{p_L},$$

yielding

$$K = \frac{\alpha p_L}{(1-\alpha)p_K}L. \tag{10.35}$$

By substituting in the first equation and solving with respect to L we deduce

$$b - \frac{\alpha p_L}{(1-\alpha)p_K}Lp_K - Lp_L = 0,$$

whence

$$L^* = \frac{b\,(1-\alpha)}{p_L}.$$

By (10.35),

$$K^* = \frac{b\alpha}{p_K}.$$

Having observed that $K^*/L^* = \alpha p_L/\,(1-\alpha)\,p_K$, from the equation

$$\lambda^* = \frac{a\,(1-\alpha)\,K^{*\alpha}L^{*-\alpha}}{p_L}$$

we also obtain

$$\lambda^* = \frac{a\,(1-\alpha)}{p_L}\left[\frac{\alpha p_L}{(1-\alpha)\,p_K}\right]^{\alpha} = a\left[\frac{\alpha}{p_K}\right]^{\alpha}\left[\frac{1-\alpha}{p_L}\right]^{1-\alpha}.$$

The Hessian of the Lagrangean is

$$L'' = \begin{bmatrix} 0 & -p_K & -p_L \\ -p_K & -a\alpha\,(1-\alpha)\,K^{\alpha-2}L^{1-\alpha} & a\alpha\,(1-\alpha)\,K^{\alpha-1}L^{-\alpha} \\ -p_L & a\alpha\,(1-\alpha)\,K^{\alpha-1}L^{-\alpha} & -\alpha a\,(1-\alpha)\,K^{\alpha}L^{-\alpha-1} \end{bmatrix}$$

with determinant

$$\det L'' = \det \begin{bmatrix} 0 & -p_K & -p_L \\ -p_K & -a\alpha\,(1-\alpha)\,K^{\alpha-2}L^{1-\alpha} & a\alpha\,(1-\alpha)\,K^{\alpha-1}L^{-\alpha} \\ -p_L & a\alpha\,(1-\alpha)\,K^{\alpha-1}L^{-\alpha} & -\alpha a\,(1-\alpha)\,K^{\alpha}L^{-\alpha-1} \end{bmatrix} =$$

$$= a\alpha\,(1-\alpha)\left(p_K^2K^{\alpha}L^{-\alpha-1} + 2p_Kp_LK^{\alpha-1}L^{-\alpha} + p_L^2K^{\alpha-2}L^{1-\alpha}\right) > 0,$$

which guarantees that we have found once more a maximum point.

10.12.3 Economic interpretation. Saddle points

We have introduced the method of Lagrange multipliers (theorem 12.1) almost without justification. The following one is very instructive and, among other things, allows us to point out the meaning of the multiplier, by solving the problem $\mathcal{P}_1$ of the *production mix* by arguments of an economic nature. Imagine that a firm wants to maximize the contribution margin $f\,(x,y)$, coming from the production/sale of two goods x and y in a given market. The constraint $g\,(x,y) = b$ can be interpreted, for example, as the maximum storage capacity of the two goods in dedicated warehouses: b could represent the square metres available in a warehouse and $g\,(x,y)$ the square metres required for the storage of the quantities x, y.

Now let us introduce a variant of $\mathcal{P}_1$, which we call $\mathcal{P}_2$ and can easily be solved. Suppose that we want to produce (x,y) in a quantity which requires fewer square metres than the capacity of the warehouse. Thus we can rent the unused part out at the current market price, let us say λ Euro for each m^2. On the contrary, if we

want to increase our production above the capacity of the warehouse, we ought to rent some space at the price λ.

In other words, assuming that we can break the warehouse constraint, two cases are possible. If $b > g(x, y)$, the contribution margin must be increased by the income $\lambda[b - g(x, y)]$ corresponding to the rent collected for the empty part of the warehouse. In the case $g(x, y) > b$, the term $\lambda[b - g(x, y)]$ has a negative sign, as it refers to an expense, and it reduces the margin $f(x, y)$. It is clear that, if λ decreases, it will probably be profitable to expand the production, while, if λ increases, it will be profitable not to use the whole warehouse and take advantage of the increase in the rent collected. The Lagrangean function

$$\boxed{L(\lambda, x, y) = f(x, y) + \lambda[b - g(x, y)]} \tag{10.36}$$

has the meaning of "adjusted contribution margin" and is a function of three variables, where one of the three, λ, is (for the time being) a constant; then we have to maximize a two-variable function without any constraint:

$$\mathcal{P}_2 : \max_{x,y} L(\lambda, x, y),$$

depending on a fixed parameter λ. We know that, if L attains a maximum at a point (x^*, y^*), its derivatives L'_x and L'_y are necessarily equal to zero at that point.

We introduce one more variant, named $\mathcal{P}_3$. Let the price λ also vary, considering that, instead of being fixed by the market in a neutral way, it may be chosen so as to damage the owner of the resource. In this case λ may have any sign. For example, if $g < b$, the firm underuses its resource and each square metre of the $b - g$ non-used resource is taxed at a "negative price λ". On the contrary, if $g > b$, the firm looks for other resources and must pay them a "positive (high) price λ".

In this new situation everything varies: λ, x, y. The hostile market chooses $\lambda = \lambda^*$ so as to minimize L, while the firm is looking for $x = x^*$ and $y = y^*$ so as to maximize L. We can formalize this third version of the problem in the following way

$$\mathcal{P}_3 : \max_{x,y} \min_{\lambda} L(\lambda, x, y).$$

There is a solution, obviously, and we just need a bit of intuition: it is sufficient to use exactly 100% of the warehouse. Then such a solution implies the respect of the constraint, that is

$$g(x, y) = b.$$

Therefore, the results concerning $\mathcal{P}_3$ are automatically transferred to $\mathcal{P}_1$. Moreover, we have identified an important kind of point:

Definition 12.1. *A point (λ^*, x^*, y^*) which minimizes L with respect to λ and maximizes L with respect to x, y is called a* ***saddle point*** *for L.*

Clearly, a saddle point is a stationary point for L. Incidentally, we note that $L(\lambda^*, x^*, y^*) = f(x^*, y^*)$. From the previous arguments we can deduce a first qualitative meaning of λ^*. Its (positive, negative or null) value represents the particular

price according to which an optimum allocation of resources *must satisfy the constraint*. Technically it is called the **shadow price**. A price $\lambda^* < 0$ means that whoever is using the resources (the firm, in our case), trying to get a maximum revenue tends to underuse them $(g < b)$, while the market drives towards their full use $(g = b)$. The opposite happens if $\lambda^* > 0$, when the firm tends to provide new resources $(g > b)$ while the market calls it to order $(g = b)$. Curious (and rare) is the case $\lambda^* = 0$: it means that the firm already attains the maximum revenue by fully using its resources, with no over- or underuse of them.

We shall return again to the economic meaning of λ^*, from a more quantitative point of view, at the end of this section. At the moment, we formalize our previous discussion in the following

Theorem 12.3. *If the triplet (λ^*, x^*, y^*) is a saddle point for the Lagrangean, then (x^*, y^*) is a solution of the problem $\mathcal{P}_1$.*

Proof. If (λ^*, x^*, y^*) is a saddle point for the Lagrangean, the following double inequality holds:

$$L(\lambda^*, x, y) \leq L(\lambda^*, x^*, y^*) \leq L(\lambda, x^*, y^*)$$

for every $\lambda \in \mathbb{R}$ and every $(x, y) \in A$, which is the domain of f and g. The right-hand inequality is equivalent to

$$(\lambda^* - \lambda)[b - g(x^*, y^*)] \leq 0$$

which implies, as it has to be true for every real number λ, that $b - g(x^*, y^*) = 0$: therefore (x^*, y^*) satisfies the constraint condition. Under this condition, the left-hand inequality becomes

$$f(x, y) + \lambda^*[b - g(x, y)] \leq f(x^*, y^*).$$

If we consider only points (x, y) satisfying the constraint, we obtain

$$f(x, y) \leq f(x^*, y^*)$$

and therefore (x^*, y^*) is a solution of $\mathcal{P}_1$. □

Meaning of λ^*

It is also possible to provide an interpretation for the numerical value of λ^*. Let us begin by noting that if b also varied, where b is the amount of the resource we considered in the constraint, the Lagrangean would become a function of four variables. In order to point out this fact, we denote the Lagrangean by the symbol $\mathcal{L}$:

$$\mathcal{L}(\lambda, x, y, b) = f(x, y) + \lambda[b - g(x, y)]. \tag{10.37}$$

We immediately note that $\mathcal{L}$ is *linear* with respect to the variable b, so that $\mathcal{L}'_b = \lambda$. Therefore, the differential of $\mathcal{L}$ is

$$\mathrm{d}\mathcal{L} = \mathcal{L}'_\lambda \mathrm{d}\lambda + \mathcal{L}'_x \mathrm{d}x + \mathcal{L}'_y \mathrm{d}y + \lambda \mathrm{d}b. \tag{10.38}$$

Let us suppose that for every b there is a stationary point (λ^*, x^*, y^*); obviously, λ^*, x^*, y^* are functions of b:

$$\lambda^* = \lambda^*(b), \quad x^* = x^*(b), \quad y^* = y^*(b),$$

as well as the optimum value $f^*(b) = f[x^*(b), y^*(b)]$. For every b, in particular, the following (stationary) conditions hold:

$$\begin{array}{l} \mathcal{L}'_\lambda[\lambda^*(b), x^*(b), y^*(b), b] = b - g[x^*(b), y^*(b)] = 0, \\ \mathcal{L}'_x[\lambda^*(b), x^*(b), y^*(b), b] = 0, \\ \mathcal{L}'_y[\lambda^*(b), x^*(b), y^*(b), b] = 0. \end{array} \tag{10.39}$$

Now we compute the differential $\mathrm{d}\mathcal{L}$ at the point $(\lambda^*(b), x^*(b), y^*(b), b)$ and we note that the partial derivatives $\mathcal{L}_\lambda$, $\mathcal{L}_x$, $\mathcal{L}_y$ are all equal to zero, according to the conditions (10.39). From (10.38) we deduce

$$\mathrm{d}\mathcal{L} = \lambda^* \mathrm{d}b. \tag{10.40}$$

We can also compute the differential $\mathrm{d}\mathcal{L}$ in another way, considering the three variables λ, x, y as depending on b; in this last case $\mathrm{d}\mathcal{L}$ measures the variation of the Lagrangean when b varies by the quantity $\mathrm{d}b$, consequently changing the other variables. If we choose this new point of view and compute again $\mathcal{L}$ at $(\lambda^*(b), x^*(b), y^*(b), b)$, from (10.37) we deduce

$$\mathcal{L}[\lambda^*(b), x^*(b), y^*(b), b] = f[x^*(b), y^*(b)] = f^*(b).$$

whence we immediately get

$$\mathrm{d}\mathcal{L} = \mathrm{d}f^*. \tag{10.41}$$

From the formulae (10.40) and (10.41) we deduce the important relationship

$$\boxed{\mathrm{d}f^* = \lambda^* \mathrm{d}b}$$

which tells us how the *shadow price* λ^* is, with a good approximation, *the ratio between the variation of the object function at the optimum point and the variation of the resource which caused it.* If we could "buy" a small additional resource on the market at the unit price p, the purchase would be profitable if $p < \lambda^*$. In the opposite case, it would be better to sell.

10.12.4 Saddle points and multipliers for n-variable functions

Let us recall the problem of the production *mix:*

$$\begin{cases} \max_{\mathbf{x}} f(\mathbf{x}) \\ sub \\ \mathbf{g}(\mathbf{x}) = \mathbf{b} \end{cases}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{b} = [b_1, b_2, \ldots, b_m]^{\mathrm{T}}$. This is an optimization problem with n variables and m constraints (with $m < n$). If f, g are differentiable,

the same arguments of the two-dimensional case with one constraint may be easily extended. Also in this case we can construct a Lagrangean, whose meaning and role are completely similar to what we have already seen:

$$L(\boldsymbol{\lambda}, \mathbf{x}) = f(\mathbf{x}) + \boldsymbol{\lambda}[\mathbf{b} - \mathbf{g}(\mathbf{x})] = f(\mathbf{x}) + \sum_{r=1}^{m} \lambda_r [b_r - g_r(\mathbf{x})].$$

In particular, theorems 12.1 and 12.3 still hold, with the trivial changes due to the higher dimension. For example:

Theorem 12.4. *If* $(\boldsymbol{\lambda}^*, \mathbf{x}^*)$, $\boldsymbol{\lambda}^* \in \mathbb{R}^m$, $\mathbf{x}^* \in \mathbb{R}^n$, *is a saddle point for* L, *then* $\mathbf{x}^*$ *is a solution of the problem* $\mathcal{P}_1$.

At $(\boldsymbol{\lambda}^*, \mathbf{x}^*)$, first order necessary conditions must be satisfied:

$$\begin{cases} \nabla_{\boldsymbol{\lambda}} L(\boldsymbol{\lambda}, \mathbf{x}) = \mathbf{b} - \mathbf{g}(\mathbf{x}) = \mathbf{0} \\ \nabla_{\mathbf{x}} L(\boldsymbol{\lambda}, \mathbf{x}) = \nabla f - \sum\limits_{r=1}^{m} \lambda_r \nabla g_r(\mathbf{x}) = \mathbf{0}. \end{cases} \tag{10.42}$$

We note that (10.42) is a system of $n+m$ equations in $n+m$ unknowns $\boldsymbol{\lambda}, \mathbf{x}$. The vector of multipliers $\boldsymbol{\lambda}^*$ is a *vector of shadow prices,* forcing the system to satisfy the constraint. Also the quantitative interpretation of $\boldsymbol{\lambda}^*$ can be extended in an almost trivial way. To each $\mathbf{b}$ we associate the stationary point $(\boldsymbol{\lambda}^*(\mathbf{b}), \mathbf{x}^*(\mathbf{b}))$. Then let $f^*(\mathbf{b}) := f[\mathbf{x}^*(\mathbf{b})]$. Thus we have

$$\mathrm{d}f^* = \boldsymbol{\lambda}^* \mathrm{d}\mathbf{b}$$

meaning, in particular, that $\dfrac{\partial f^*}{\partial b_j} = \lambda_j^*$ for every $j = 1, ..., m$.

10.13 Exercises

10.1. Determine the domain of the following functions. Decide whether they are open, closed or bounded sets.

$$f(x,y) = \sqrt{\frac{x-y}{x+2y}}, \quad g(x,y) = \ln\left(4 - x^2 - y^2\right) - \ln x.$$

10.2. Draw the level curves of the following functions in the x, y-plane

$$f(x,y) = \frac{x+y}{x-y}, \quad g(x,y) = \left(x^2+y^2\right) e^{-\left(x^2+y^2\right)}, \quad h(x,y) = \frac{1+xy}{x^2}.$$

10.3. Write the matrices associated with the following quadratic forms and classify them.

(*a*) $q(x_1, x_2) = x_1^2 + 3x_2^2 - 2x_1x_2$,
(*b*) $q(x_1, x_2) = 2x_1^2 - x_2^2 + x_1x_2$,
(*c*) $q(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_1x_2 + 3x_1x_3 - 6x_2x_3$,
(*d*) $q(x_1, x_2, x_3) = x_1^2 + 2x_2^2 + 3x_3^2 + x_1x_2 + 3x_1x_3$.

10.4. For each of the following functions, compute the first partial derivatives and the first differential.

$$(a)\ \ f(x,y)=3x^3e^{-xy},\quad (b)\ \ g(x,y,z)=z^2\ln(x-y).$$

For the first function, write the equation of the tangent plane at the point $(1,0)$ and compute the second derivatives and the second differential at $(1,0)$.

10.5. Let

$$Y(K,L)=a\left[dK^{-\rho}+(1-d)L^{-\rho}\right]^{-1/\rho}\qquad 0<d<1,\ \rho>0.$$

Compute the partial derivatives of Y.

10.6. Let

$$F(x,y)=\frac{1+iy}{1+ix},\quad G(x,y)=(1+i)^{y-x},\quad i>0.$$

Compute $\mathrm{D}_y\ln F(x,y)$ and $\mathrm{D}_y\ln G(x,y)$, and verify that they do not depend on x.

10.7. Write the second order Maclaurin's formula for the functions

$$(a)\ \ f(x,y)=x-y+3x^2-5xy-6yx^{2001},\quad (b)\ \ g(x,y)=\ln(1+x)e^y.$$

10.8. Let $Q=\sqrt{LK}$. Verify that

$$Q=L\frac{\partial Q}{\partial L}+K\frac{\partial Q}{\partial K}.$$

10.9. In a neighbourhood of the point $(1,1)$, the equation

$$x\ln y+y\ln x=0$$

implicitly defines

(a) $y=f(x)$, (b) $x=g(y)$, (c) both $y=f(x)$ and $x=g(y)$

(d) neither $y=f(x)$ nor $x=g(y)$.

10.10. The costs (C) of a firm depend on the number of employees (S) and the physical volume of advertising (A), according to the formula

$$C(S,A)=0,7S^3A^{0,5}.$$

(a) Compute the differential of C.

(b) By means of the differential, determine the percentage variation of costs due to a $2,5\%$ growth in the number of employees and to a 0,8% reduction in the volume of advertising.

10.11. Suppose that among the costs for the firm mentioned in the previous exercise, besides employees and advertising, we also take into account inventory costs (I) according to the formula

$$C(S,A)=0,7S^3A^{0,5}I^{0,5}.$$

Evaluate the effect of a 4% variation of the inventory costs in both of the following cases:

(a) the two other factors remain unchanged,

(b) the two other factors vary as in the previous exercise.

10.12. Let

$$Q = L^{\alpha} K^{1-\alpha}$$

be a production function and

$$L^{\alpha} K^{1-\alpha} = c, \qquad c > 0$$

one of its level curves (*isoquantum*). Compute the slope $\dfrac{\mathrm{d}K}{\mathrm{d}L}$ on the isoquantum[14].

10.13. Determine the stationary points of the following functions and decide their nature.

(a) $f(x,y) = x^3 + y^3 - 3xy$,

(b) $f(x,y) = x^4 + y^3 - 4x^2 - 3y^2$,

(c) $f(x_1,x_2) = \left(x_1^2 + x_2^2\right) e^{-\left(x_1^2+x_2^2\right)}$,

(d) $f(x_1,x_2) = 5\ln x_1 + 2\ln x_2 - 0,1x_1 - 0,4x_2$,

(e) $f(x_1,x_2,x_3) = \dfrac{1}{x_1} + \dfrac{1}{x_2} + \dfrac{1}{x_3} + x_1x_2x_3$,

(f) $f(x,y,z) = x^2 - 2x + y^2 + \ln\left(1+z^2\right)$.

10.14. A firm is studying the use of two kinds of chemical products in the physical quantities x and y. The amount z of product obtained - to be maximized - depends on the quantities of the two chemicals used, according to the statistically estimated relation:

$$z = f(x,y) = 1000\ln(1+x) + 3000\ln(1+y) - 2x - 3y.$$

Determine the maximum point of the function f.

10.15. Determine and decide the nature of all the extrema of the following functions, where the variables are subject to the constraint indicated beside the function.

(a) $f(x,y) = xy, \qquad x^2 + y^2 = 2$,

(b) $f(x,y) = \ln(x-y), \quad x^2 + y^2 = 2$ and $x > y$.

Are they local or global constrained extrema?

10.16. Determine all possible extrema of

$$f(x,y) = -x\ln x - y\ln y$$

under the condition

$$x + y = 1.$$

---

[14] The slope $\dfrac{\mathrm{d}K}{\mathrm{d}L}$ on an isoquantum has a picturesque name: *marginal rate of technical substitution* ($MRTS$), and represents the decreasing (why not increasing?) rate of the capital, due to a unit increment of labour, which leaves the production volume unchanged.

10.17. Verify that, given any Cobb-Douglas production function $Q(K, L)$ and referring to the problem

$$\begin{cases} \max_{K,L} Q(K, L) \\ sub \\ Kp_K + Lp_L = b, \end{cases}$$

at the maximum point we have

$$\frac{\partial Q/\partial K}{\partial Q/\partial L} = \frac{p_K}{p_L}.$$

10.18. A garage offers a car and lorry service, with a production function

$$Q = 15K^{1/3}L^{2/3}.$$

The unit cost of labour is 6 Euro, the cost of capital is 3 Euro.

(a) Use the method of Lagrange multipliers in order to maximize the production, with a forecast budget of $b = 450$ Euro. Decide whether, with such costs, one may think of increasing the budget in order to attain a 10% increase of the maximum of the production.

(b) Optimize the budget, forecasting a production $Q = 10000$.

10.19. A firm must decide how many pages x and y of advertising ought to be purchased in two magazines, in order to promote a new product. Each page costs $p_1 = 1$ in the first magazine and $p_2 = 2$ in the second. The firm aims at maximizing the volume of sales

$$S(x, y) = ax + by + cxy,$$

with $a, b, c > 0$, spending the allocated amount $B = 5$. Determine, with the method of Lagrange multipliers, the maximum point of the function S.

11

Financial Calculus

Financial Calculus is interesting for its own sake, as the importance of financial activities in contemporary society has rapidly increased; and it also features many important applications of the topics of a standard Mathematics course.

We present here the basics of this Calculus. Our main aim is to help students in understanding the examples which are scattered throughout this book. It is not really a course in Financial Mathematics, as this would require more pages and a more adequate development. Let us call it a survival kit.

The main features of this chapter are as follows.

- We introduce the two processes of *accumulation* and *discount*.
- This leads us to the study of the most common *financial laws*.
- We then deal with some *practical applications*:

— the value of annuities;
— the amortization of debts;
— the comparison of the profitability of various financial opportunities.

11.1 Accumulation and discount

Basic vocabulary

Financial calculus deals with the exchange of amounts of money which refer to different maturities.

Let us suppose we buy a bond today. If we buy a zero coupon bond (*zcb* from now on) and pay 1000 Euro for it, and we know that in 3 months we will receive 1030 Euro, we are actually exchanging the two kinds of goods "Euro today" and "Euro at a future maturity" (more precisely: "Euro at the maturity 3 months").

The exchange rate is

$$\frac{1030}{1000} = 1.03$$

This operation is called *accumulation*, and we can think of it as transferring money forwards in time.

The invested amount is generally denoted by P and called the *principal*; the received amount is generally denoted by A and named the *accumulated amount*. The quotient $f := A/P$ is called the *accumulation factor*. The difference $I := A - P$ is called *interest*.

In the example above we have $P = 1000$, $A = 1030$, $f = 1.03$ and $I = 30$.

Let us now consider the same operation from the point of view of the company issuing the bond (or of the Government, in the case of State bonds). The issuer is bound to pay 1030 Euro in 3 months' time, but it immediately receives 1000 Euro. Once again we have an exchange of "Euro today" against "Euro at a future maturity", but the issuer will think of it as an operation of transferring money backwards in time, from the maturity 3 months to today. This kind of operation is called *discount*.

The future amount is generally denoted by N and named the *future value,* or *nominal value*; the present amount is generally denoted by P and named the *present value*. The quotient $\phi := P/N$ is called the *discount factor*. The difference $D := N - P$ is called *discount*.

In our example $N = 1030$, $P = 1000$, $\phi = \frac{1000}{1030} = 0.97087$ and $D = 30$.

Basic relations

The basic relation between principal and accumulated amount follows from the definition of the accumulation factor and is given by

$$A = Pf.$$

We also frequently use the additive relation:

$$A = P + I.$$

Generally speaking, the multiplicative relation $A = Pf$ is more useful. However, the two relations refer to equivalent representations and the link between them is obvious:

$$I = P\,(f - 1) \text{ and } f = 1 + \frac{I}{P}.$$

In discount problems, similarly, the basic relation between N and P is

$$P = N\phi.$$

Alternatively, we can use the relation

$$P = N - D.$$

The link between them is given by

$$D = N(1-\phi) \text{ and } \phi = 1 - \frac{D}{N}.$$

As accumulation and discount are symmetric exchange operations, the corresponding exchange ratios f, ϕ are linked. If we apply to an amount P, over the same period of time, first the accumulation which takes it forwards in time and then the discount which takes it backwards in time, we should again find the same amount P:

$$(Pf)\phi = P.$$

Therefore the two factors f and ϕ are linked by the following relations

$$f\phi = 1 \; ; \quad \phi = \frac{1}{f} \; ; \quad f = \frac{1}{\phi}.$$

Two factors f, ϕ such that their product is equal to one are called *conjugate factors.* In general, accumulation factors and discount factors are called *financial factors.*

As we shall see, in financial practice we start by defining a rule of calculation for constructing an accumulation factor f or a discount factor ϕ. Then the notion of conjugate factor allows us to find, from the rule of calculation which regards one kind of financial factor (for example: the accumulation factor), the other financial factor (for example: the discount factor) which is coherent with it.

Of course, these factors will depend on many circumstances and situations. Typically, they are functions of time: namely, of the time interval which represents the period for which we invest our money (in accumulation) or the distance in time of a future amount of money (in discount). Therefore, we write $f(t)$, $\phi(t)$ to recall this fact.

The way time acts on the value of each financial factor is governed by a parameter, generally called the interest rate or discount rate, which sets the rate at which the passing of time influences the growth of the accumulation factor - or the increasing distance in time of a future amount of money influences its value today. We will write $f(t,\alpha)$, $\phi(t,\beta)$ when we want to underline the fact that the financial factors also depend on these parameters.

In many practical situations it is better to separate the two possible approaches. Therefore, we will use:

— *a single function* of time $f(t)$ or $\phi(t)$,

— or *an infinity of functions* of time $f(t,\alpha)$ or $\phi(t,\beta)$, corresponding to the infinite values the above parameters can take.

Definition 1.1. *A function $f(t,\alpha)$ or $\phi(t,\beta)$, depending on time and on a parameter, determines a* ***system of financial laws****, respectively accumulation laws or discount laws. When we fix the value of the parameter we get a single function of time $f(t)$ or $\phi(t)$, and therefore we get a single* ***financial law,*** *respectively an accumulation law or a discount law.*

We will see that the infinite system of functions of time $f(t,\alpha) = 1 + \alpha t$ describes the *system of laws of simple interests.* If we fix $\alpha = 10\%$ we get the single function $f(t) = 1 + 0.1t$, which describes the law of simple interests at 10%.

We will now give a more precise meaning to the notions of interest rate and discount rate, which we introduced above in an intuitive way.

Definition 1.2. *The interest produced by one Euro, invested for one year, is called the (annual)* ***interest rate****:*

$$\textit{interest rate} = i := f(1) - 1.$$

The amount of money retained as a compensation by someone who anticipates today the amount of one Euro, which is due in one year, is called the (annual) ***discount rate***:

$$\textit{discount rate} = d := 1 - \phi(1).$$

Similar notions can be introduced with a different choice of the unit of measure for time. We can therefore have semiannual rates, monthly rates,... However, in financial calculations the most commonly used unit for time is the year.

If we have two conjugate financial factors the relation $f(t)\phi(t) = 1$ must hold for all values of t; therefore, (annual) interest and discount rates must satisfy the relation:

$$(1+i)(1-d) = 1,$$

and solving this with respect to i or to d we obtain:

$$d = \frac{i}{1+i}, \qquad i = \frac{d}{1-d}.$$

From a practical point of view, the interest rate is the *percentage* increment in the first year of an invested principal, while the discount rate is the *percentage* reduction of a future amount of money which is anticipated for one year.

We note that an interest rate is applied for one specific kind of goods: "Euro today", while a discount rate is applied for another kind of goods: "Euro in one year's time". These are two different kinds of goods, therefore the two kinds of rates are conceptually different.

Finally, we remark that the relations $A = Pf$ and $P = N\phi$ establish a *proportionality relation* between the accumulated amount and the principal, and between the present value and the nominal value. Obviously, the same proportionality holds between interest and principal and between discount and nominal value.

11.2 Standard systems of financial laws

There exist three *standard systems of financial laws*:

— simple interests (and simple discount);

— compound interests (and compound discount);

— bank discount (and anticipated simple interests).

We shall now consider the most relevant formulae for each of them. The most natural way of introducing them is by means of an accumulation factor f for the first two, and of a discount factor ϕ for the third one.

11.2.1 Simple interest and simple discount

Simple interest is defined by the fact that the amount of interests is proportional (not only to the principal P but also) to the time interval t which defines the accumulation process:

$$I \text{ proportional to } Pt$$

therefore

$$\frac{I}{Pt} = \text{ constant.}$$

Now, when $P = 1$ and $t = 1$ we have $I = i$ (interest rate); this means that the value of the constant must be exactly i. From $I/Pt = i$, we get

$$I = Pit,$$

and finally

$$A = P + I = P + Pit = P \cdot (1 + it)$$

For simple interest, therefore, $f(t) = 1 + it$.

For example, if $P = 1000$, $i = 12\% = 0.12$, $t = 3/12 = 3$ months $= 0.25$ years, the accumulated amount is

$$1000 \left(1 + 12\% \cdot \frac{3}{12}\right) = 1030. \tag{11.1}$$

Definition 2.1. *The accumulation factor*

$$f(t) = 1 + it$$

defines the system of ***simple interest****. The conjugate discount factor*

$$\phi(t) = \frac{1}{1 + it}$$

defines the system of ***simple discount*** *(also: rational discount). The rate i is called the (annual) simple interest rate.*

What happens if we move from the unit of measure "year" to another unit - let us say to the month? Is it possible to define a monthly interest rate i_{12} (the index 12 reminds us that 12 months make one year) which leads us to the same accumulation factor, when we express time in months? The answer is rather obvious: t years will become $12t$ months, therefore i_{12} has to satisfy the relation

$$1 + it \equiv 1 + i_{12} \cdot 12t,$$

whence

$$i_{12} = \frac{i}{12}.$$

In general, the interest rate i_m corresponding to the fraction $1/m$ of a year is given by

$$i_m = \frac{i}{m}.$$

The interest rate i_m is called a *period interest rate*. The interest rates i and i_m are said to be *equivalent rates* (*in simple interest*).

For example, the quarterly interest rate which is equivalent to the annual rate 12% is

$$i_4 = \frac{12\%}{4} = 3\%.$$

If we use this to compute the accumulated amount corresponding to the same investment we dealt with just before relation (11.1), we find the encouraging result:

$$1000\,(1 + 3\% \cdot 1) = 1030.$$

An application of simple interests: zcb and BOT

The Italian bonds called *Buoni Ordinari del Tesoro* (BOT) are a good example of zero-coupon bonds (zcb). In Italy, calculations on BOT are commonly carried out using simple interest; in the US these kinds of bonds are called T-Bills, and calculations on them are commonly performed using bank discount (see below).

Let us briefly examine a zcb. In everyday practice, only the purchase price P and the reimbursement value N (greater than P) of a bond with maturity T years after the purchase are given, and they completely define the investment. This kind of bond does not pay coupons at intermediate maturities, therefore in a way it never explicitly pays interests. Initially, zero-coupon bonds became very popular because they fitted very well into a flaw of the US taxation system: as no payment of interests formally took place, there could be no taxation. This fact made the operation very convenient both for the issuers (as they could guarantee better conditions) and for the subscribers (because of the tax exemption), notwithstanding the fact that the difference $N - P$ did actually represent - then and now - interest paid to the subscribers. Nowadays, taxation systems can "see" these interests and tax them. However, zcb are still issued and exchanged, mainly because of their very simple structure. They are the elementary bricks with which modern financial engineering builds many complicated financial tools.

We have seen the real meaning of the difference $N - P$; let us now try to define what simple interest rate r can describe the return on our investment.

For a zcb, the simple interest rate such that the accumulated amount corresponding to the purchase price is equal to the reimbursement value is called the *simple yield to maturity*. In formulae, we have

$$P\,(1 + rT) = N,$$

whence

$$r = \frac{N - P}{PT}.$$

We note that the purchase price of the zcb is also the present value - using simple discount - of the reimbursement value, when the rate is given by the simple yield:

$$P = \frac{N}{1 + rT}.$$

If we resell the bond t years later (with $t < T$), at the price P' which is determined by the current simple yield r' in the financial market at that date, we must have:

$$P' = \frac{N}{1 + r'(T - t)},$$

As we only held the bond for the time period going from 0 to t, the corresponding simple yield must be

$$r'' = \frac{P' - P}{Pt} = \frac{\dfrac{N}{1 + r'(T - t)} - \dfrac{N}{1 + rT}}{t\dfrac{N}{1 + rT}},$$

whence

$$r'' = \frac{rT - r'(T - t)}{t[1 + r'(T - t)]}.$$

We note that if the current simple yield r' in the market at date t is exactly coincident with the simple yield to maturity r, the simple yield r'' we really get through buying the bond today and selling it at date t is *not equal to* r - it is smaller:

$$r'' = \frac{rT - r(T - t)}{t[1 + r(T - t)]} = \frac{r}{1 + r(T - t)} < r.$$

11.2.2 Compound interests and compound discount

We lend a principal of one Euro for one year, at the interest rate i. After one year, our credit becomes

$$1 + i.$$

Let us suppose the debtor does not pay us anything, not even the interest, and we decide to be patient. This actually means we increase the loan we grant to the debtor to $1 + i$ Euro, for the following year. Therefore, after one more year, the debtor owes us the amount of $1 + i$ Euro plus the interest $i(1 + i)$ on that amount. That is, our credit becomes

$$1 + i + i(1 + i) = (1 + i)^2.$$

In general, after n years our credit becomes

$$(1 + i)^n.$$

If the time period t of the loan is not an integer, we can generalize again and introduce the accumulation factor

$$f(t) = (1 + i)^t.$$

The process we described above, through which the interest accrued in a certain period of time becomes part of the principal and then - starting from the following period of time - produces more interest, is called *compound interest*.

Definition 2.2. *The accumulation factor*

$$f(t) = (1+i)^t$$

defines the system of ***compound interest****. The conjugate discount factor*

$$\phi(t) = \frac{1}{(1+i)^t} = (1+i)^{-t}$$

defines the system of ***compound discount.*** *The rate i is called the (annual) compound interest rate.*

For example, if we invest a principal $P = 1000$ Euro for two years at compound interest, with the annual interest rate $i = 10\%$, the accumulated amount becomes

$$A = Pf = P(1+i)^2 = 1000 \cdot (1+10\%)^2 = 1210.$$

In a similar way, the present value of an amount $N = 1000$ computed 6 months earlier at compound interest, with the annual interest rate $i = 10\%$, is

$$P = N\phi = 1000 \cdot \frac{1}{(1+10\%)^{1/2}} = 953.46.$$

In compound interests, the relation between equivalent rates is more complex. The period interest rate i_m corresponding to $1/m$ year must satisfy

$$(1+i_m)^{mt} = (1+i)^t \text{ for all } t, \tag{11.2}$$

where

$$f(t) = (1+i_m)^{mt}$$

is the expression of the accumulation factor in terms of the period interest rate.

The relation (11.2), in turn, holds if and only if the following formula holds:

$$(1+i_m)^m = 1+i, \text{ that is } i_m = (1+i)^{1/m} - 1,\ i = (1+i_m)^m - 1. \tag{11.3}$$

We shall say, for example, that in compound interests the annual rate $i = 21\%$ and the semi-annual rate $i_2 = 10\%$ are *equivalent rates*, as they generate the same accumulation factors. In fact, they satisfy the formulae (11.3) for $m = 2$:

$$10\% = (1+21\%)^{1/2} - 1, \qquad 21\% = (1+10\%)^2 - 1.$$

We have seen how we can describe a family of financial factors f or ϕ in terms of annual interest rates i or period interest rates i_m. In practice *two other ways* are used to describe this family:

— through nominal interest rates,

— through instant interest rates.

We shall consider instant rates later; but let us introduce nominal rates now.

If we start again with the period rate i_m corresponding to $1/m$ year, we call the quotient between the period rate i_m and the width of the time period $1/m$ the *nominal annual rate*. We shall denote it by j_m:

$$j_m = \frac{i_m}{1/m} = m i_m.$$

The rate j_m depends on m; this leads us to say that the interest rate j_m is *convertible m times per year* or - which is the same - every $1/m$ year. For example, in the case of a semi-annual rate $i_2 = 10\%$ we obtain the nominal annual rate which is convertible twice a year

$$j_2 = 2 \times 10\% = 20\%.$$

We note that given a period rate i_m the equivalent annual rate i gives precisely the same interest produced by the period rate i_m on a principal of one Euro in one year, while the nominal annual rate j_m always gives a little less than that. For this reason, we frequently say that i is an *effective annual rate*.

As $i_m = j_m/m$, we can also write the expression of the accumulation factor $f(t)$ in terms of nominal rates

$$f(t) = \left(1 + \frac{j_m}{m}\right)^{mt}.$$

11.2.3 Bank discount and anticipated simple interests

An individual has a credit of amount N, which is due in t years. He transfers it to a bank, exchanging it with the immediate payment of an amount $P < N$. The difference $D = N - P$ is calculated by the bank to be proportional both to N and to the lifespan t of this anticipated amount:

$$D \text{ proportional to } Nt$$

and therefore

$$\frac{D}{Nt} = \text{constant}.$$

If the future value N is one Euro and the amount is advanced for one year, we have $D = d$ (discount rate); this means that the value of the constant is d. Therefore

$$D = Ntd,$$

whence

$$P = N - Ntd = N(1 - td).$$

The factor $\phi = 1 - td$ is a discount factor.

Definition 2.3. *The discount factor*

$$\phi(t) = 1 - td$$

defines the system of ***bank discount***. *The conjugate accumulation factor*

$$f(t) = \frac{1}{1 - td}$$

defines the system of ***anticipated simple interests****. The rate d is called the (annual) bank discount rate.*

We note that if we consider the same operation we described above from the point of view of the bank which acquires the credit, it is an investment. The name of anticipated simple interests which is commonly used to describe the related accumulation factor is justified by the fact that the bank first of all withholds the amount $D = Ntd$, as a compensation for the money being advanced, and this amount is computed exactly as with simple interests (the product of the amount of money, times time, times rate).

The lifespan t of the operation and the discount rate d must satisfy the condition $td < 1$, otherwise the value of the transferred credit would be null or even negative!

When this system is used for a lifespan t which is near to $t^* = 1/d$, the product td approaches 1 and this produces some odd consequences. Let us think of investing 1000 Euro for 5 years at the bank discount rate $d = 18\%$. The accumulated amount becomes:

$$1000\frac{1}{1 - 18\% \cdot 5} = 10000,$$

and this means that the value of the investment has increased tenfold. If the lifespan of the investment is prolonged for 6 months, from 5 to 5.5 years, the accumulated amount would become:

$$1000\frac{1}{1 - 18\% \cdot 5,5} = 100000.$$

The value of the investment has now increased a hundredfold, and this is clearly unacceptable - especially for the individual who has to pay the accumulated amount.

For this system of financial laws we do not say anything about period rates, as they are rarely used. We just mention that the formulae are similar to those of simple interests.

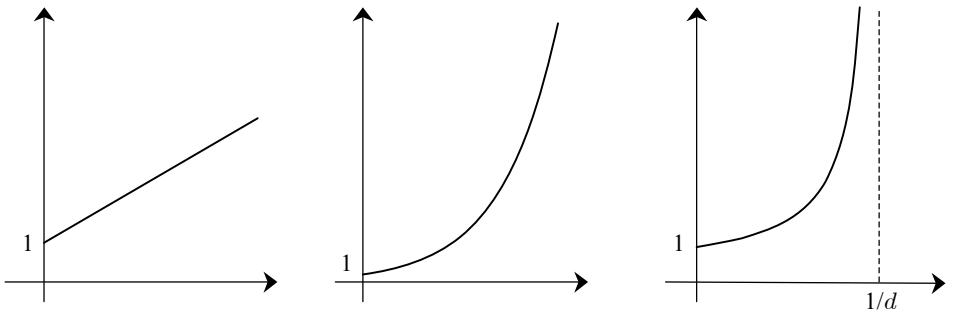


Figure 11.1. Graphs of the accumulation factors in the three standard financial systems

Final values and present values, with several amounts of money

In practice we shall frequently deal with a list of amounts of money $a_1, a_2, \ldots, a_n$, with maturities $t_1, t_2, \ldots, t_n$. We shall say that these amounts of money are a *cash flow*.

The amounts of money will be considered as positive if they refer to amounts we receive, and as negative if they refer to amounts we have to pay. Frequently, the list of amounts we have to consider is made up of amounts with all the same sign (either all positive, or all negative). In this case, we shall say that they are an *annuity*.

The single amounts are called *instalments* or terms of the annuity, and we usually want to compute:

- the present value of the annuity, that is the global present value of all the given amounts at the date 0, which comes before all the given maturities, or

- the accumulated amount (more often called: final value) of the annuity, that is the global accumulated amount of all the given amounts at a date T, which comes after all the given maturities.

We meet the first problem, for example, if a firm wants to transfer a portfolio of credits with different maturities, and looks for the present value which is the correct compensation for this transfer. We meet the second problem, for example, when an investment assures certain amounts of money at certain maturities, and we want to know what is the final value the investment will guarantee at a date T if we profitably reinvest all those amounts till date T.

Obviously, the present value P of an annuity will be given by the sum of the present values of all its terms, computed with discount factors $\phi(t_s)$ depending on the maturities:

$$P = \sum_{s=1}^{n} a_s \phi(t_s)$$

and similarly, the final value F will be given by the sum of the accumulated amounts of all its terms, computed with accumulation factors $f(T - t_s)$ depending on the length of the time intervals between maturities t_s and the final date T:

$$F = \sum_{s=1}^{n} a_s f(T - t_s).$$

Example 2.1. Consider the annuity given by the following terms at the corresponding maturities:

Maturity	**Instalment**
1	1000
2	2500
4	3000

Its present value, in simple discount, with annual interest rate 4%, is

$$P = \frac{1000}{1 + 4\%} + \frac{2500}{1 + 4\% \cdot 2} + \frac{3000}{1 + 4\% \cdot 4} \simeq 5862.6$$

while its final value, at simple interest, computed at date 6, that is two years after the last maturity, with annual interest rate 5%, is

$$F = 1000\,[1 + 5\% \cdot (6-1)] + 2500\,[1 + 5\% \cdot (6-2)] + \\ +3000\,[1 + 5\% \cdot (6-4)] = 7550.$$

We shall see below that in some special cases it is not really necessary to add up all the present values, or all the accumulated amounts, thanks to some useful formulae.

11.2.4 Force of interest

An important way of describing how the accumulated amount of an investment is constituted leads us to define the notion of force of interest.

We invest one Euro for t years and call the accumulated amount we obtain $f(t)$. At that date we could interrupt our investment and cash the accumulated amount. We wonder whether this option is more advantageous than continuing our investment for h more years (for example $h = 1/365$, that is one day). In this case we would gain more interests, for the amount:

$$f(t+h) - f(t).$$

Now what is the simple interest rate r, such that investing $f(t)$ at the interest rate r for h years would lead us to the same result? The answer is quite simple: we just need to solve the following equation with respect to r:

$$f(t)\,rh = f(t+h) - f(t),$$

as it expresses the fact that the interests we obtain in the two ways are equal. We get:

$$r = \frac{f(t+h) - f(t)}{f(t)\,h}. \tag{11.4}$$

This rate r is a function of both t and h. From (11.4), if the function f is differentiable at t we can write:

$$r = \frac{f'(t)}{f(t)} + \frac{o(h)}{h}.$$

Therefore, if h is small enough the rate r is correctly approximated by the function

$$\rho(t) := \frac{f'(t)}{f(t)} = \mathrm{D}\left[\ln f(t)\right]$$

Definition 2.4. *The function*

$$\rho(t) = \frac{f'(t)}{f(t)}$$

is called the ***force of interest*** *corresponding to the accumulation factor* $f(t)$.

Let us now consider the three standard financial systems and compute the force of interest for each of them.

Simple interests. The force of interest is

$$\rho(t) = \frac{f'(t)}{f(t)} = \frac{i}{1+it}$$

where i is the annual interest rate.

Let us check the meaning of $\rho(t)$, through a numerical example. Consider the rate $10\% = 0.1$:

$$f(t) = 1 + 0.1t \tag{11.5}$$

Take $t = 2$. At date 2 the accumulated amount of 1 Euro is:

$$f(2) = 1.2.$$

If we continue our investment for one day more, the additional interest will be:

$$f\left(2 + \frac{1}{365}\right) - f(2) = 0.1 \cdot \frac{1}{365} = 0.00027397. \tag{11.6}$$

The force of interest at 2 is:

$$\rho(2) = \frac{0.1}{1 + 0.1 \cdot 2} = 0.08\overline{3}.$$

The approximation to the correct amount of additional interest (11.6) obtained through the force of interest is:

$$f(2) \cdot \rho(2) \cdot \frac{1}{365} = 1.2 \cdot 0.08\overline{3} \cdot \frac{1}{365} = 0.00027397$$

which is very good (as it is correct up to the eighth decimal digit). The force of interest $\rho(2) = 8.\overline{3}\%$ simply tells us that if we continue our investment for a short period of time after two years, with the same accumulation factor defined by (11.5), this is equivalent to investing the amount we accumulated in two years at the simple interest rate $8.\overline{3}\%$.

Compound interests. The force of interest is:

$$\rho(t) = \mathrm{D}\left[\ln f(t)\right] = \mathrm{D}\left[\ln(1+i)^t\right] = \mathrm{D}\left[t\ln(1+i)\right] = \ln(1+i).$$

This time we obtained a constant; in compound interests (and this is the only case!) the force of interest does not actually depend on time t.

The symbol δ is often used for it in this case, and δ is also called the *instant interest rate* for compound interests.

Suppose we fix the value of the nominal annual interest rate j_m to be $j_m = \delta$. Let us see what happens if the accumulation of interests becomes more and more

frequent (that is, if m continues to increase). As we know, for each value of m the accumulation factor is

$$f(t) = \left(1 + \frac{\delta}{m}\right)^{tm}$$

As we have already seen in Chapter 3 on sequences, when m diverges the accumulation factor $f(t)$ will approach the factor

$$f(t) = e^{\delta t}.$$

We can check this statement by computing $f(1) = (1 + \delta/m)^m$ with $\delta = 10\%$, for some values of m.

With a semi-annual accumulation ($m = 2$) we have $\left(1 + \frac{10\%}{2}\right)^2 = 1.1025$, with a monthly accumulation we have $\left(1 + \frac{10\%}{12}\right)^{12} = 1.1047$, with a daily accumulation $\left(1 + \frac{10\%}{365}\right)^{365} = 1.1052$, and this is the result we expected as $e^{10\% \cdot 1} = 1.1052$.

Finally, the link between the force of interest δ and the annual interest rate i comes from the requirement that they generate the same accumulation factors:

$$(1 + i)^t = e^{\delta t} \text{ for all } t,$$

and this gives again

$$\delta = \ln(1 + i), \text{ and } i = e^{\delta} - 1.$$

For example, the force of interest which is equivalent to the annual interest rate $i = 20\%$ is $\delta = \ln(1 + 20\%) = 18.232\%$.

Anticipated simple interests. The force of interest is:

$$\rho(t) = \mathrm{D}[\ln f(t)] = \mathrm{D}\left[\ln \frac{1}{1 - dt}\right] = \frac{d}{1 - dt}.$$

The three expressions we obtained for the force of interest behave very differently in time:

— in simple interests, $\rho(t) = i/(1 + it)$ tends to 0 as t diverges;

— in compound interests, $\rho(t) \equiv \ln(1 + i)$ remains a constant;

— in anticipated simple interests, $\rho(t) = d/(1 - dt)$ diverges as t approaches $1/d$ (from the left).

It is useful to understand why this happens, and try to increase our awareness of the mechanisms of financial calculus.

We recall that $\rho(t)$ represents the simple interest rate which, when it is applied to the accumulated amount $f(t)$ already obtained at t for an additional short period of time h, reproduces with a good approximation the interests $f(t + h) - f(t)$ that $f(t)$ determines in the same period h. That is:

$$f(t + h) - f(t) \simeq f(t) \cdot \rho(t) \cdot h, \text{ when } h \text{ is small.}$$

In the case of *simple interests*, f produces

$$f(t+h) - f(t) = [1 + i(t+h)] - [1 + it] = ih,$$

which is constant with respect to time t because for simple interests only the principal we started with can produce an interest. The force of interest is the quotient between this additional interest and the product of the amount $f(t)$ at time t and the additional period of time h

$$\rho(t) = \frac{ih}{(1+it)h} = \frac{i}{1+it},$$

therefore it has to decrease, as the additional interest does not increase in time while $f(t)$ does increase.

In the case of *compound interests*, f produces an interest which increases in time proportionally to $f(t)$ itself:

$$\begin{aligned} f(t+h) - f(t) &= (1+i)^{t+h} - (1+i)^t = \\ &= (1+i)^t \left[(1+i)^h - 1\right] \simeq (1+i)^t \cdot \ln(1+i) \cdot h. \end{aligned}$$

(we all remember notable limits, don't we?). The fact that this additional interest is proportional to $f(t) = (1+1)^t$, together with the definition of the force of interest, immediately give us the fact that $\rho(t)$ has to be structurally constant.

In the case of *anticipated simple interests*, the explosive speed with which the interest increases when we approach the forbidden date $t^* = 1/d$ is such that $\rho(t)$ diverges to infinity.

11.3 Typical applications of compound interests

11.3.1 Simple annuities with constant instalments

We say that an annuity is *simple* if the period between two consecutive maturities is always the same, and this is also equal to the period of time to which the interest rate refers (we note that the word "simple" here has no relation with simple interests). In our examples below, this period will always be equal to one year.

Let us consider a simple annuity, which is *unitary* (that is: we get one Euro a year), for n years. Let us suppose that payments take place at the end of each year. Let us compute the final value at the end of the last year, and the present value at the beginning of the first year, using compound interest at the annual rate i. The final value is computed exactly when we get the last payment (that is, at year n), and the present value one year before the first payment (that is, at year 0).

The final value is traditionally denoted by the symbol $s_{\overline{n}|i}$, which reads "s corner n at rate i" (or "s angle n at rate i", or "s figured n at rate i") and its value is

$$s_{\overline{n}|i} = (1+i)^{n-1} + (1+i)^{n-2} + \cdots + 1 = \frac{(1+i)^n - 1}{i},$$

as we can easily check, using the formula for the sum of terms in a geometric progression we saw in Chapter 1.

The present value is traditionally denoted by the symbol $a_{\overline{n}|i}$, which reads "a corner n at rate i" (or "a angle n at rate i", or "a figured n at rate i"), and using the same formula we get

$$a_{\overline{n}|i} = \frac{1}{1+i} + \frac{1}{(1+i)^2} + \cdots + \frac{1}{(1+i)^n} = \frac{1-(1+i)^{-n}}{i}.$$

If our annuity is not unitary, and we get R Euro a year, we just multiply the results by R; therefore the present value is $P = Ra_{\overline{n}|i}$ and the final value is $F = Rs_{\overline{n}|i}$.

Example 3.1. Compute the final value and the present value of an annuity of 1000 Euro a year for 10 years, at compound interest, with annual rate 5%. We get

$$s_{\overline{10}|5\%} = \frac{1.05^{10}-1}{0.05} = 12.578$$

and

$$F = 1000 \cdot 12.578 = 12578;$$

then we get

$$a_{\overline{10}|5\%} = \frac{1-1.05^{-10}}{0.05} = 7.7217$$

and

$$A = 1000 \cdot 7.7217 = 7721.70.$$

If the annuity is not *ordinary* as above (that is: instalments are paid at the end of each year), but *due* (instalments are paid at the beginning of each year), the present value is calculated when we get the first payment (that is, at year 0) and the final value is calculated one year after the last payment (that is, at year n).

The symbols for due annuities are even uglier: $\ddot{a}_{\overline{n}|i}$ and $\ddot{s}_{\overline{n}|i}$, and they can be read "$a$ due corner n at rate i" and "s due corner n at rate i". Their expressions can be obtained using the same procedure as above (we leave all computations to the reader):

$$\ddot{a}_{\overline{n}|i} = (1+i)\,\frac{1-(1+i)^{-n}}{i} \qquad \text{and} \qquad \ddot{s}_{\overline{n}|i} = (1+i)\,\frac{(1+i)^n-1}{i}.$$

We add that the two present values $a_{\overline{n}|i}$ and $\ddot{a}_{\overline{n}|i}$ can also be computed for *simple perpetuities* (annuities with an infinite number of instalments). The corresponding symbols are $a_{\overline{\infty}|i}$ and $\ddot{a}_{\overline{\infty}|i}$, and their values are:

$$a_{\overline{\infty}|i} = \frac{1}{i} \qquad \text{and} \qquad \ddot{a}_{\overline{\infty}|i} = (1+i)\,\frac{1}{i},$$

as we can see by taking the expressions related to a finite number of instalments and letting n diverge to $+\infty$. Otherwise, we can also obtain them directly as sums of geometric series (see Chapter 6).

Finally, if the period between two consecutive maturities is always the same but it is not a year (six months, or one month, or whatever) we can use the same formulae as above, provided we use the correct unit of measure for time and the correct interest rate.

Example 3.2. We get 1000 Euro every six months for 5 years, and we want to compute the final value F, at compound interest, at the annual rate 21%. We obtain first of all the semi-annual rate

$$i_2 = (1 + 21\%)^{1/2} - 1 = 10\%.$$

Then we remark that 5 years correspond to 10 semesters (six-month periods), therefore

$$F = 1000 s_{\overline{10}|10\%} = 1000 \cdot \frac{1.10^{10} - 1}{0.1} = 15937.$$

If we also want the present value P, we have:

$$P = 1000 a_{\overline{10}|10\%} = 1000 \cdot \frac{1 - 1.1^{-10}}{0.1} = 6144.60.$$

11.3.2 Discounted Cash Flow

Let us consider a general financial operation (for example: the purchase of a bond). The operation is fully described by its cash flow, that is a list of amounts of money $a_0, a_1, \ldots, a_n$ with maturities $t_0, t_1, \ldots, t_n$. As we have said, positive amounts refer to amounts we receive and negative amounts refer to amounts we pay. It is often convenient to accept also null amounts, meaning that at a certain maturity there is no money flow. Finally, we shall usually choose the origin of the time-line as our first maturity: $t_0 = 0$.

The *Discounted Cash Flow* (DCF) of a financial operation is the algebraic sum of the present values of its flows, computed at $t_0 = 0$ using compound discount. This algebraic sum is seen as a function $G(x)$ of the interest rate x we use in compound discount. In general, therefore:

$$G(x) = a_0 + a_1 (1 + x)^{-t_1} + a_2 (1 + x)^{-t_2} + \cdots + a_n (1 + x)^{-t_n}.$$

For example, if our flows are:

Maturities	**Flows**
0	−100000
1	90000
2	−40000

the DCF is

$$G(x) = -100000 + \frac{90000}{1 + x} - \frac{40000}{(1 + x)^2}.$$

What is the use of the DCF? In financial analysis, it is useful in at least two cases.

(a) We are interested in the DCF of a financial operation, computed at a certain interest rate i, as this number represents the financial value of the operation for all

individuals who usually invest their money at rate i (this rate is called the *opportunity cost of capital*).

Definition 3.1. *The number $G(i)$ is called the* ***Net Present Value*** *(NPV) of the financial operation, computed at the interest rate i.*

(b) We are interested in all interest rates $x^* > -1$ for which the *DCF* is null:

$$G(x^*) = 0,$$

because in a sense they are a measure of the return on an investment, or of the cost of a loan.

We note that till now we have always supposed an interest rate i to satisfy the condition $i \geq 0$. In this case, however, we also consider rates i such that $-1 < i < 0$.

When $i = -1/2 = -50\%$, for example, this means that if we invest one Euro for t years the accumulated amount is $f(t) = (1 - 1/2)^t = (1/2)^t < 1$. Therefore we do not really get interests, and we lose a part of our principal. When $i = -1 = -100\%$, this means that if we invest one Euro for t years the accumulated amount is $f(t) = (1-1)^t = 0$. Therefore we do not get interest and we lose all our principal; this could happen, actually, but we would have a serious problem when discounting as the denominator $1 + x$ would be null. If the rate i is such that $i < -1$ we would have even more serious problems. For example, if $i = -1.5 = -150\% < -1$ and the investment period is 6 months = 0.5 years, the result of the investment would be $f(t) = (1 - 1.5)^{1/2} = \sqrt{-0.5}$, which is clearly meaningless.

Definition 3.2. *An interest rate $x^* > -1$ for which the* DCF *of a financial operation is null is called an* ***internal rate*** *(or implicit rate) for the operation.*

More specifically, we call it the *Internal Rate of Return* in the case of investments and the *Internal Rate of Cost* in the case of loans; for more general financial operations, which are neither pure investments nor pure loans, we just call these interest rates (if they exist) internal rates.

Let us try and understand the reason for these names, at least for investments. If we invest one Euro for 2 years at 10% (annual compound), the accumulated amount is 1.21. The DCF of the operation is

$$G(x) = -1 + \frac{1.21}{(1+x)^2},$$

which is null at $x^* = 10\%$, therefore confirming that our return is exactly 10%.

If we invest one Euro for 2 years, and every year we receive simple interest at 12%, after one year we get 0.12 and after two years 1.12 (because the invested principal also has to be returned). The DCF of the operation is

$$G(x) = -1 + \frac{0.12}{1+x} + \frac{1.12}{(1+x)^2},$$

and it is null at $x^* = 12\%$, therefore telling us what is the return on our investment.

In each of these cases the equation $G(x^*) = 0$ can be rewritten as

$$\text{initial investment} = \text{discounted value of future incomes}$$

confirming the fact that in some way x^* describes the effective conditions of the investment.

Suppose now that we invest one Euro today, in one year we get 0.11 as interest and in two years we get back not only the principal and 0.11 Euro as interest, but also a *reimbursement premium* of 0.025 Euro. What is the return on this new investment? If we say it is 11%, that is the mere quotient between interests and invested principal ($0.11/1 = 0.11$; this value is actually called the *immediate return* on the investment), we are now underestimating the profitability of the investment; but how can we take into account the reimbursement premium, which characterizes this not completely standard operation? Here is where DCF plays an important role. In the first two cases the DCF had proved to be reliable: it was null exactly in correspondence with the "natural" return the investments gave. Let us trust it again, and see when it is null in this case. The DCF is now

$$G(x) = -1 + \frac{0.11}{1+x} + \frac{1 + 0.025 + 0.11}{(1+x)^2}$$

and the corresponding equation is

$$\frac{0.11}{1+x^*} + \frac{1.135}{(1+x^*)^2} = 1$$

The solution is now $x^* \simeq 12.178\%$, and we could take this as a reliable measure of the return on the investment.

We note that we should be careful when using this information. Is it better to buy a bond whose return is 12.178% or a bond whose return is 12%? Intuitively we would say the former, but maybe we should be more cautious. It is very different whether we get a return of 12.178% with a *zero-coupon bond*, without reinvestment problems along the way, or whether we get the same return with a bond which pays us coupons along the way: in this case, to understand which is the better choice we must also take into account the reinvestment conditions we get for our intermediate incomes.

The next example will tell us something more about this kind of financial decision.

Example 3.3. Let us suppose an individual wants to invest money at the conditions we described at the beginning of this paragraph. Suppose that she usually invests money at 4%. If we compute the NPV of the (not very advantageous) operation she is about to subscribe to, using the 4% rate, we get:

$$G(4\%) = -100000 + \frac{90000}{1.04} - \frac{40000}{1.04^2} = -50444.$$

This means that investing money in that operation is equivalent to her losing *now* 50444 Euro. To see this even better, suppose she has exactly 100000 Euro; let us study the situation she would face in two years, according to two different scenarios.

(a) She immediately loses 50444 Euro and invests what she has left at 4%. In this case she remains today with $100000 - 50444 = 49556$ Euro, invests this sum at 4% for two years, which then amounts to:

$$49556 \cdot 1.04^2 = 53600.$$

(b) She subscribes to the financial operation. In this case, after one year she gets 90000 Euro. She invests this at 4%, as usual, and after one year this amounts to $90000 \cdot 1.04 = 93600$. At that date, she has to pay 40000 Euro. Therefore she is left with:

$$93600 - 40000 = 53600.$$

After two years, the situation is exactly the same in the two cases.

It now appears very natural to accept the idea that, if we want to compare several financial operations in order to choose the most advantageous one, it is sufficient to compare the NPVs of those operations using our opportunity cost of capital as the evaluation rate: the best operation is the one which features the highest NPV.

Let us return to our financial operation. We try to find its internal rate, to have an idea of the conditions at which the investor has invested her money. The equation we want to solve is

$$-100000 + \frac{90000}{1+x} - \frac{40000}{(1+x)^2} = 0$$

but unfortunately it has no solutions. This operation has no internal rate.

Let us see what happens, if at year 2 we replace the outflow of 40000 Euro with an income of 80000 Euro. The situation gets much better. In fact the NPV becomes

$$G(4\%) = -100000 + \frac{90000}{1.04} + \frac{80000}{1.04^2} = 60503,$$

therefore investing money in the operation is equivalent to getting now 60503 Euro. The internal rates are found as solutions of the equation

$$-100000 + \frac{90000}{1+x} + \frac{80000}{(1+x)^2} = 0.$$

and the only acceptable solution is $x^* = -\frac{11}{20} + \frac{1}{20}\sqrt{401} \simeq 45.125\%$.

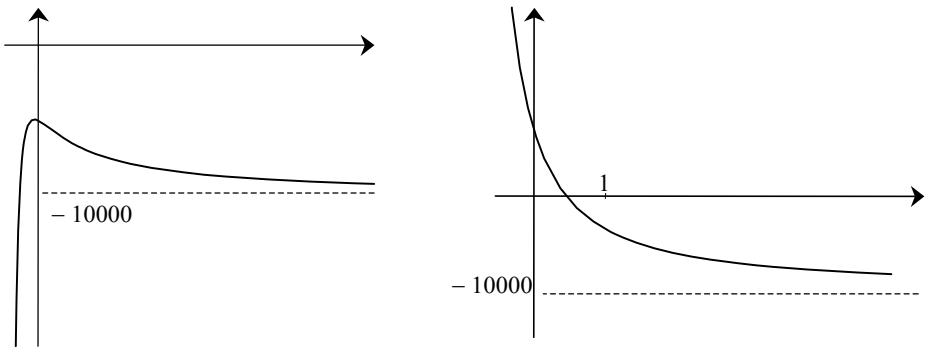


Figure 11.2. The graphs of DCF in the two operations

We can see that, at least in this case, both the positive sign of the NPV and the value of the internal rate tell us that the investment is advantageous. In the case

of the first operation, however, the negative sign of the NPV correctly told us that the operation was not advantageous while there was no internal rate - and therefore the internal rate could give us no information. We cannot get any further with this question, but in Financial Mathematics we learn that the NPV gives a criterion for decisions which is better than the criterion given by the internal rate.

11.3.3 Amortization plans

A firm gets a loan of amount S from a bank at date 0, and has to pay it back by returning to the bank the amounts of money $R_1, R_2, \ldots, R_n$, called *amortization instalments*, at maturities $t_1 < t_2 < \cdots < t_n$.

For many reasons, we are interested in analysing the operation of repayment of the loan in some detail. We analyse how the initial debt $D_0 = S$ reduces in time, assuming that the value is D_s at intermediate dates t_s till it vanishes at t_n where $D_n = 0$. We also determine which part of each instalment is meant to be paid as interest on the current amount of the debt and which part is meant to be paid as a mere repayment of the capital debt S itself: we therefore speak of *interest shares* and *principal shares* (or *capital shares*).

The DCF of the operation for the bank is

$$G(x) = -S + \frac{R_1}{(1+x)^{t_1}} + \frac{R_2}{(1+x)^{t_2}} + \cdots + \frac{R_n}{(1+x)^{t_n}}$$

The interest rate of the loan is the internal rate x^* which is the solution of equation $G(x^*) = 0$. In general, we cannot find an exact expression for it but we can approximate it numerically quite easily.

Through this rate we can build the sequence $D_0, D_1, D_2, \ldots, D_n$ of *residual debts* at maturities $0, t_1, t_2, \ldots, t_n$. It is enough to consider the fact that the debt D_s at a given maturity t_s is obtained from the debt D_{s-1} at the preceding maturity t_{s-1}, taking into account the fact that we first increase it by taking it forward from t_{s-1} to t_s and then we have to subtract the instalment R_s we pay at time t_s:

$$D_s = D_{s-1}(1+x^*)^{t_s - t_{s-1}} - R_s. \tag{11.7}$$

The increase is calculated at the interest rate x^* we just considered. This relation is rather intuitive and very useful.

Example 3.4. The amount of a loan is $S = 1000$ Euro and we can repay it in two years, with two annual payments of amounts $R_1 = 700$ and $R_2 = 440$ respectively. The DCF for the bank which is financing the loan is

$$G(x) = -1000 + \frac{700}{1+x} + \frac{440}{(1+x)^2}.$$

Its only acceptable solution is $x^* = 0,1 = 10\%$. The sequence of residual debts is

$$\begin{cases} D_0 = 1000 \\ D_1 = D_0 \cdot (1 + 10\%) - 700 = 400 \\ D_2 = D_1 \cdot (1 + 10\%) - 440 = 0. \end{cases}$$

The principal shares after 1 and 2 years are

$$\begin{cases} C_1 = D_0 - D_1 = 600 \\ C_2 = D_1 - D_2 = 400. \end{cases}$$

The interest shares are obtained from the instalments, subtracting the principal shares:

$$\begin{cases} I_1 = 700 - 600 = 100 \\ I_2 = 440 - 400 = 40. \end{cases}$$

We get the *amortization plan*:

Maturity	Instalment	Principal share	Interest share	Residual debt
0	-	-	-	1000
1	700	600	100	400
2	440	400	40	0

11.3.4 A theoretical issue: decomposability

Let us assume we invest our financial means for $s + t$ years at compound interest. We can write the accumulation factor in the following way:

$$(1+i)^{s+t} = (1+i)^s (1+i)^t, \tag{11.8}$$

and this tells us that the result does not vary if instead of a single investment for $s+t$ years we consider an investment for s years, immediately followed by a reinvestment for t years at the same conditions.

More generally, if we call f a general accumulation factor, we can write a condition for the final result to be independent of possible interruptions in the investment, as follows:

$$f(s+t) = f(s) f(t). \tag{11.9}$$

Definition 3.3. *If an accumulation law f is such that for all $s, t \geq 0$ the relation (11.9) holds, we say it is* ***decomposable*** *(also:* ***separable****).*

Elementary properties of powers guarantee (see relation (11.8)) that compound interests define a decomposable accumulation law.

It is quite natural to ask whether there are other accumulation laws f with the same property. The answer is negative, as we have already seen in Chapter 4; exponential functions are the only continuous functions for which (11.9) holds:

If $f(s+t) = f(s) f(t)$ for all $s, t \geq 0$, then there exists $i > -1$ such that:

$$f(t) = (1+i)^t.$$

We can conclude that decomposability (also: separability) is a property which holds only for the laws of compound interests.

In practice, we often use only three systems of financial laws (simple interests, compound interests, anticipated simple interests); therefore many students think that

compound interests is the only decomposable system among these three, while with other systems of financial laws decomposability may still be an open question. Please note: that is wrong! The function $f(t) = (1+i)^t$ is the one and only mathematical function $f(t)$ which has this property.

11.4 Exercises

12.1. Compute the accumulated amounts using simple interest, compound interest and anticipated simple interest of a principal $P = 1000$ Euro, at rate 10%, after 3 years. The rate is an interest rate in the first two cases and a discount rate in the third one.

12.2. Discount the amount of 1000 Euro for 3 years using simple discount, compound discount and bank discount, at rate 10%. The rate is an interest rate in the first two cases and a discount rate in the third one.

12.3. An individual has rights to receive 1000 Euro at the end of the year for 5 years and 1200 Euro at the end of the year for 5 more years. Compute the present value of such an annuity, using compound discount at the interest rate 5%.

12.4. A debt of 1000 Euro is repaid in two years, with two payments at the end of each year of amount 700 and 600 respectively. Find the internal rate x^* of the loan, and find the residual debt after one year.

12.5. Consider the law of simple interests with interest rate s and the law of anticipated simple interests with discount rate d. Suppose $s > d$, and find the positive maturity t^* for which the two financial laws give the same accumulated amount.

12.6. An individual can invest using simple interests at rate $s > 0$. As an alternative, he can invest using compound interests at rate $c > 0$. For which values of c is the accumulated amount with compound interest larger than the other, regardless of the maturity $t > 0$?

12.7. At the maturities 0,1,2 a financial operation originates the cash flow -1000, r, 800. Find r such that the internal rate of the operation is 40%. For that value of r, compute the NPV at rate 30%. Draw the graph of the DCF of the operation, for rates $x > -1$.

12.8. What is the relation between the quarterly interest rate i_4 and the force of interest δ, at compound interests, if the two rates are equivalent rates?

12.9. From the relations $\delta = \ln(1+i)$ and $i = e^\delta - 1$ which give the equivalence between annual rate and force of interest at compound interest, deduce approximate polynomial relations between the two rates using Taylor's formula to the second order.

12.10. An individual invests the amount P for n years, and at the end of each year she gets simple interests at the annual rate r. She reinvests these amounts of interest at compound interest, at the annual rate i, till the final maturity n. This operation then involves a principal P at year 0 and an accumulated amount A (which includes the repayment of the principal) at year n. Find the corresponding accumulation

factor f, which is a function of r and i: $f(r,i)$. What happens in the special case $r = i$?

12.11. An individual buys a car which costs 4000 Euro. He pays 1000 Euro now, and has to pay 18 monthly instalments of 200 Euro each at the end of each month. At the beginning he also has to pay 150 Euro for loan expenses, and on each instalment he has to pay an additional 1% for collection expenses. Compute the actual cash flow of the operation at all maturities. What is the total amount of interests (all expenses included) related to the loan we described above? Write the DCF $G(x)$ of the operation, from the point of view of the bank which issues the loan. Without actually computing the internal rate of the operation (all expenses included), can you tell if it is more or less than 20%? (The internal rate of a loan, all expenses included, is called *TAEG - Tasso Annuo Effettivo Globale* - in Italy).

Index

MyBook http://mybook.egeaonline.it

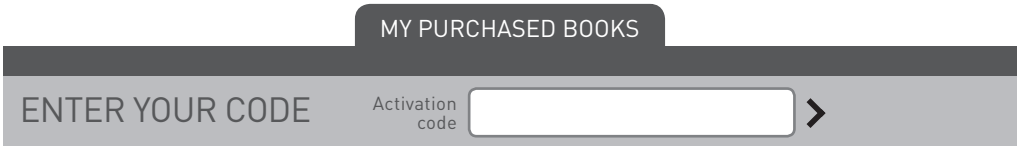
Type your username and password or register by clicking on **Create a new account**.

In MyBook you can access the accompanying resources (both text and multimedia), the **BookRoom**, the **EasyBook** app and your purchased books.

CODE

5939JaBX12

Type the code in the **Activation Code** field.



The code must be typed only the first time you access **MyBook** and cannot be used thereafter.

This textbook is tailored for those educational programs, such as Economics and Management, which include a first (and frequently the only) course of Mathematics. We have selected some topics which we consider to be fundamental, if not mandatory, for these students:

- the knowledge of Calculus, for functions of one and two variables;
- the use of Calculus in optimization;
- the notion of integral for functions of one variable;
- the language and the elementary techniques of Linear Algebra;
- the basics of Financial Calculus.

Several preliminary examples from applied sciences (mainly from Economics) introduce the theoretical aspects. We have tried to avoid an excessive formalism, in order to quickly reach the fundamental concepts.

Lorenzo Peccati is Senior Full Professor of Mathematics at the Università Bocconi, Milan, Italy.
Sandro Salsa is Full Professor of Mathematical Analysis at the Politecnico di Milano, Milan, Italy.
Annamaria Squellati was formerly Lecturer of Mathematics at the Università Bocconi, Milan, Italy.