



Mariano Giaquinta
Giuseppe Modica

Mathematical Analysis

Approximation and Discrete Processes

Birkhäuser



An engraving from Mario Bettini *Aerarium philosophiae mathematicae*, 1648. The prince provides money with which the Society of Jesus educates people in mathematics for *aesthetic* and *practical* purposes.

Mariano Giaquinta
Giuseppe Modica

Mathematical Analysis
Approximation and Discrete Processes



Birkhäuser
Boston • Basel • Berlin

Mariano Giaquinta
Scuola Normale Superiore
Dipartimento di Matematica
I-56100 Pisa
Italy

Giuseppe Modica
Università degli Studi di Firenze
Dipartimento di Matematica Applicata
I-50139 Firenze
Italy

Library of Congress Cataloging-in-Publication Data

Giaquinta, Mariano, 1947-

[Analisi matematica. 2, Approssimazione e processi discreti. English]

Mathematical analysis : approximation and discrete processes / Mariano Giaquinta,

Giuseppe Modica.

p. cm.

Includes bibliographical references and index.

ISBN 0-8176-4313-3 (alk. paper)

1. Mathematical analysis. I. Modica, Giuseppe. II. Title.

QA300.G49713 2004

515—dc22

2004043696

CIP

AMS Subject Classifications: 00A35, 01A20, 01A35, 01A40, 01A45, 05-01, 11-01, 26-01, 26A03, 26A15, 26A16, 26A18, 26A45, 26B05, 30B10, 34-01, 37-01, 40-01, 41-01, 60-01

ISBN 0-8176-4313-3

Printed on acid-free paper.

Printed on acid-free paper.
©2004 Birkhäuser Boston

Birkhäuser 

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Birkhäuser Boston, c/o Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to property rights.

Printed in the United States of America. (TXQ/HP)

9 8 7 6 5 4 3 2 1

SPIN 10890261

Birkhäuser is a part of *Springer Science+Business Media*

www.birkhauser.com

Preface

This volume¹ aims at introducing some basic ideas for studying *approximation processes* and, more generally, *discrete processes*. The study of discrete processes, which has grown together with the study of infinitesimal calculus, has become more and more relevant with the use of computers. The volume is suitably divided in two parts.

In the first part we illustrate the numerical systems of *reals*, of *integers* as a subset of the reals, and of *complex numbers*. In this context we introduce, in Chapter 2, the notion of *sequence* which invites also a rethinking of the notions of limit and continuity² in terms of discrete processes; then, in Chapter 3, we discuss some elements of *combinatorial calculus* and the mathematical notion of *infinity*. In Chapter 4 we introduce *complex numbers* and illustrate some of their applications to elementary geometry; in Chapter 5 we prove the *fundamental theorem of algebra* and present some of the elementary properties of polynomials and rational functions, and of finite sums of harmonic motions.

In the second part we deal with discrete processes, first with the process of *infinite summation*, in the numerical case, i.e., in the case of *numerical series* in Chapter 6, and in the case of *power series* in Chapter 7. The last chapter provides an introduction to *discrete dynamical systems*; it should be regarded as an invitation to further study.

We have tried to keep the treatment of topics as independent as possible even at the cost of some repetition; usually, we assume as known the content of [GM1], but, whenever possible, we provide an alternative elementary treatment in order to allow the use of part of this volume on sequences and series, independently from infinitesimal calculus.

The main body is formed by Chapter 1, Sections 2 and 3, Chapter 2, Sections 1, 2, 3, and 4, Chapter 4, Sections 1 and 2, Chapter 6, Sections 1, 2, 3, and 4 and Chapter 7, Sections 1 and 2 for about a third of the whole. The rest of the material may appear as heterogeneous; it develops in branches that eventually meet, from which it is easy to select several paths. However,

¹ This volume is a translation and revised edition of M. Giaquinta, G. Modica, *Analisi Matematica, II, Approssimazione e processi discreti*, Pitagora Editrice, Bologna, 1999.

² We have discussed these notions in M. Giaquinta, G. Modica, *Mathematical Analysis. Functions of One Variable*, Birkhäuser, Boston, 2003. In this volume we shall refer to this work as [GM1].

we believe that the whole of the material is, besides its intrinsic interest, fundamentally basic for any further study of mathematical analysis.

As in [GM1] an appropriate number of exercises are distributed in the text and at the end of each chapter. They are marked by the symbol ¶; the double ¶¶ indicates exercises that are more difficult.

We are greatly indebted to Cecilia Conti for her help in polishing our first draft and we warmly thank her. We would like to thank also Alessandro Berarducci, Roberto Conti, Pietro Majer and Stefano Marmi for their comments when preparing the Italian edition, and Stefan Hildebrandt for his comments and suggestions concerning especially the choice of illustrations. Our special thanks go also to all members of the editorial technical staff of Birkhäuser for the excellent quality of their work and especially to the executive editor Ann Kostant.

Note: We have tried to avoid misprints and errors. But, as most authors, we are imperfect authors. We will be very grateful to anybody who wants to inform us about errors or just misprints or wants to express criticism or other comments. Our e-mail addresses are

`giaquinta@sns.it`

`modica@dma.unifi.it`

We shall try to keep up an errata corrige at the following webpage:

<http://www.sns.it/~giaquinta>

Mariano Giaquinta
Giuseppe Modica
Pisa and Firenze
October 2003

Contents

Preface	v
1. Real Numbers and Natural Numbers	1
1.1 Introduction	1
a. Numbers and measurement	2
b. Never-ending processes	4
c. Back to numbers	6
d. An axiomatic or a constructive approach?	8
1.2 The Axiomatic Approach to Real Numbers	9
1.2.1 Algebraic and order properties	9
a. Axioms for addition	10
b. Axioms for multiplication	10
c. The distributive law	11
d. Order	12
1.2.2 Continuity property	13
a. Supremum	13
b. The extended real line	15
c. Dedekind cuts of \mathbb{R}	15
1.2.3 Uniqueness of reals	16
1.3 Natural Numbers	17
a. Natural numbers and the principle of induction ..	17
b. Approximation of reals by rational numbers	20
c. Recursive statements	21
1.4 Summing Up	25
1.5 Exercises	26
2. Sequences of Real Numbers	31
2.1 Sequences	31
a. Limit of a sequence	35
b. Properties of limits and calculus	36
c. Limits of monotone sequences	39
d. Sequences and supremum	40
e. Subsequences	40
2.2 Equivalent Formulations of the Continuity Axiom	41
a. The principle of nested intervals or Cantor's principle	41

	b. Cauchy criterion	42
	c. Upper and lower limits	44
	d. Bolzano–Weierstrass theorem	46
	e. The continuity property of the reals	46
2.3	Limits of Sequences and Continuity	47
	a. Limits of sequences and limits of functions	47
	b. Continuity in terms of sequences	48
2.4	Some Special Sequences	49
	a. Elementary limits	50
	b. Powers, exponentials and factorials	53
	c. Wallis and Stirling formulas	55
	d. Numerical integration	57
2.5	An Alternative Definition of Exponentials and Logarithms ..	59
	a. A definition of a^x using continuity	59
	b. Euler’s number e	62
	c. Derivative of the exponential	62
2.6	Summing Up	63
2.7	Exercises	65
3.	Integer Numbers: Congruences, Counting and Infinity ..	71
3.1	Congruences	71
	3.1.1 Euclid’s algorithm	71
	a. The greatest common divisor	72
	b. Integer solutions of first order equations	74
	3.1.2 Prime factorization	77
	3.1.3 Linear congruences	79
	3.1.4 Euler’s function ϕ	82
	3.1.5 RSA Cryptography	84
3.2	Combinatorics	88
	3.2.1 Samples, mappings and subsets	89
	a. Ordered samples and mappings	89
	b. Nonordered samples and subsets	91
	c. Ordered lists	92
	d. The formula of inclusion and exclusion	93
	e. Surjective maps	94
	3.2.2 Drawings	95
	3.2.3 Location problems	95
	3.2.4 The hypergeometric and multinomial distributions ..	97
3.3	Infinity	99
	3.3.1 The mathematical analysis of infinity	99
	a. Cardinality	100
	b. Cantor–Bernstein theorem	102
	c. Denumerable sets	103
	d. The axiom of choice	104
	e. The power of the continuum	105
	f. The continuum hypothesis	106
	3.3.2 Some information on the theory of sets	107

3.4	Summing Up	111
3.5	Exercises	113
4.	Complex Numbers	121
4.1	Complex Numbers	122
	a. The system of complex numbers	122
	b. The n -th roots	128
	c. Complex exponential and logarithm	129
4.2	Sequences of Complex Numbers	131
	a. Definitions	131
	b. Weierstrass's theorem	132
4.3	Some Elementary Applications	133
	4.3.1 A few applications of the complex notation	133
	4.3.2 A few applicatons to elementary Euclidean geometry	135
	a. Special points of a triangle	136
	b. Equilateral triangles	138
4.4	Summing Up	140
4.5	Exercises	142
5.	Polynomials, Rational Functions and Trigonometric Polynomials	145
5.1	Polynomials	145
	5.1.1 The Division Algorithm	147
	a. Euclid's algorithm and Bezout identity	148
	b. Factorization	150
	c. The factor theorem	150
	5.1.2 The fundamental theorem of algebra	153
	a. Factorization in \mathbb{C}	153
	b. Simple and multiple roots of a polynomial	156
	c. Factorization in \mathbb{R}	157
5.2	Solutions of Polynomial Equations	158
	5.2.1 Solutions by radicals	158
	5.2.2 Distribution of the roots of a polynomial	163
	a. Descartes's law of signs	164
	b. Sturm's theorem	164
5.3	Rational Functions	166
	a. Decomposition in \mathbb{C}	166
	b. Decomposition in \mathbb{R}	170
	c. Integration of rational functions	171
5.4	Sinusoidal Functions and Their Sums	173
	5.4.1 Trigonometric polynomials	174
	a. Periodic functions	174
	b. Trigonometric polynomials	175
	c. Spectrum and energy identity	176
	d. Sampling	178
	5.4.2 Sums of sinusoidal functions	181
5.5	Summing Up	183

5.6	Exercises	185
6.	Series	187
6.1	Basic Facts	188
	a. Definitions and examples	189
	b. A necessary condition for convergence	192
	c. Series and improper integrals	192
	d. Decimals	193
6.2	Taylor Series, e and π	195
	a. The number π	198
	b. More on the number e	202
6.3	Series of Nonnegative Terms	204
	a. Series of positive decreasing terms	206
	b. The root and ratio tests	209
	c. Viète's formula for π	211
	d. Euler and Wallis formulas	212
6.4	Series of Terms of Arbitrary Sign	214
	a. Absolute convergence	214
	b. Series of complex terms	215
6.5	Series of Products	216
	a. Alternating series	216
	b. Summation by parts	219
	c. Sequences of bounded total variation	219
	d. Dirichlet and Abel theorems	221
6.6	Products of Series	222
6.7	Rearrangements	225
6.8	Summing Up	227
6.9	Exercises	230
7.	Power Series	235
7.1	Basic Theory	238
	7.1.1 Circle of convergence	238
	a. The disc and the domain of convergence	240
	7.1.2 Continuity of the sum	241
	a. Uniform Convergence	241
	b. Continuity of uniform limits	242
	c. Uniform convergence of power series	243
	7.1.3 Differentiation and integration	244
	a. Series of derivatives and of integrals	244
	b. Real power series	245
	c. Power series and Taylor series	247
	d. Complex series	248
7.2	Further Results	250
	7.2.1 Boundary values	250
	7.2.2 Product and composition of power series	253
	a. Weierstrass's double series theorem	254
	7.2.3 Taylor series: examples	254

7.3	Some Applications	257
7.3.1	Complex functions	258
7.3.2	An alternate definition of π , e and of elementary functions	260
7.3.3	Series solutions of differential equations	262
7.3.4	Generating functions and combinatorics	264
a.	Generating functions	264
b.	Enumerators	266
c.	Exponential enumerators	268
d.	A few location problems	269
e.	Partitions of a set	271
7.4	Further Applications	274
7.4.1	Euler–MacLaurin summation formula	274
a.	Bernoulli numbers	274
b.	Bernoulli polynomials	276
c.	Euler–MacLaurin formula and Stirling’s approximation	278
7.4.2	Euler Γ function	280
a.	Definition and characterizations	280
b.	Functional relations	282
c.	Asymptotics of Γ and ψ	286
7.5	Summing Up	288
7.6	Exercises	289
8.	Discrete Processes	297
8.1	Recurrences	302
8.1.1	Linear difference equations	302
a.	First order linear difference equations	302
b.	Second order homogeneous difference equations	304
c.	Second order nonhomogeneous difference equations	305
d.	\mathcal{Z} -transform and Laplace transform	306
e.	Fibonacci’s numbers	308
8.1.2	Some nonlinear examples	311
a.	Simple examples	311
b.	Evaluating algorithm performance	312
c.	Rate of convergence	314
8.1.3	Continued fractions	316
a.	Definitions and elementary properties	316
b.	Developments as continuous fractions	321
c.	Infinite continued fractions	323
d.	Irrationals and approximations by rationals	326
e.	Order of approximation and transcendental numbers	329
8.2	One-Dimensional Dynamical Systems	331
8.2.1	Discretization and models	332
a.	Euler’s method	332

b. Runge-Kutta method	334
c. Models	335
8.2.2 Examples of one-dimensional dynamics	337
a. Expansive dynamics	338
b. Contractive dynamics: fixed points	338
c. Sinks and sources	339
d. Periodic orbits	340
e. Periodic-doubling cascade transition to chaos	342
f. The intermittency phenomenon	344
g. Ergodic dynamics	345
8.2.3 Chaotic dynamics	348
a. Sensitive dependence on initial conditions and the Lyapunov exponent	349
b. Chaotic orbits	351
c. Bernoulli's shift	351
d. The triangular map	353
e. Conjugate maps	354
8.2.4 Chaotic attractors, basins of attraction	354
8.2.5 Cantor sets and other self-similar sets	357
a. Measure and dimension	357
b. Cantor sets	360
c. Iterated function systems	362
d. Dimension of the invariant set	363
8.3 Two-Dimensional Dynamical Systems	369
8.3.1 Game of life	369
8.3.2 Fractal boundaries	370
a. Julia sets	371
b. Mandelbrot set	372
8.3.3 Fractals on the computer	372
8.4 Exercises	373
A. Mathematicians and Other Scientists	377
B. Bibliographical Notes	379
C. Index	381

1. Real Numbers and Natural Numbers

In this chapter, after an introductory section, in Section 1.2 we shall illustrate the axiomatic approach to real numbers, and, in Section 1.3, we shall identify the natural numbers as the *smallest inductive subset of \mathbb{R}* . Further information about natural numbers will be discussed in Chapter 3, while the notions of *sequences* and of *limit of a sequence*, which are specially relevant in mathematics, are discussed in Chapter 2; in Section 2.2 we present, in particular, several equivalent formulations of the *continuity axiom*.

1.1 Introduction

Rudiments of mathematics, or even refined geometrical and algebraic rules appear in many ancient civilizations, as for instance the Babylonian, the Egyptian, the Hindu, the Chinese or some of the pre-Colombian civilizations. But mathematics as an organized, independent and reasoned discipline, that is as a science, developed from 600 to 300BC in Greece, thanks probably to the democratic political system of the Greeks that must have encouraged the attitude toward arguing.

Thales of Miletus (624BC–546BC) is given credit for inventing the *mathematical proof*, and, according to Diadochus Proclus (411–485), Pythagoras of Samos (580BC–520BC)

changed the study of geometry into the form of a liberal education, for he examined the principles to the bottom, and investigated its theorems in an immaterial and intellectual manner.

Most of our sources are, however, of several centuries later and refer to Thales and Pythagoras in a legendary and mythological way. For example Aristotle, reporting on the mystic-religious society of Pythagoreans, says:

the so-called Pythagoreans applied themselves to the study of mathematics . . . ; in so much that, having been brought up in it, they thought that its principles must be the principles of all existing things. . . . They thought they found in numbers more than in fire, earth, or water, many resemblances to things which

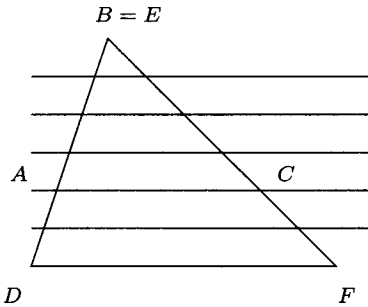


Figure 1.1. Thales's theorem.

*are and become Since then, all other things seemed in their whole nature to be assimilated to numbers, while numbers seemed to be the first things in the whole of nature, they supposed the elements of numbers to be the elements of all things, and the whole heaven to be a musical scale and a number.*¹

a. Numbers and measurement

It seems therefore that the Pythagoreans believed all bodies to be made up of a great number of corpuscles that were all identical and harmoniously arranged. They identified integer numbers with patterns of those atoms, thus making integers the basis of measure. Geometrical entities, such as lines, surfaces and solids existed, as any other aspect of reality, as aggregations of point-numbers. This is probably why they came to the conclusion that the relations of capacity between two homogeneous quantities could always be evaluated in terms of the ratio of positive integer numbers, by counting in principle the number of corpuscles in the quantities. Concluding the argument, two homogeneous quantities seem to be always *commensurable*.

From this point of view the process of measurement becomes that of

- (i) finding (with a finite procedure) a unit of measure e , possibly the largest, common to the quantities to be measured,
- (ii) counting; if the quantity A is n -times e , and a quantity B is m -times e , then the relation between A and B is expressed by the quotient of the integers n and m .

In fact the basic proofs of some geometrically relevant facts seem to be in favour of the assumption of commensurability. Here are a few examples.²

1.1 Theorem (Thales's theorem). *Let ABC and DEF be two triangles with equal angles. If the segments AB and DE are commensurable*

¹ Aristotle (384BC–322BC), *Metaphysics*.

² This actually only shows that some geometric constructions preserve *rationality* of the measures of the data.

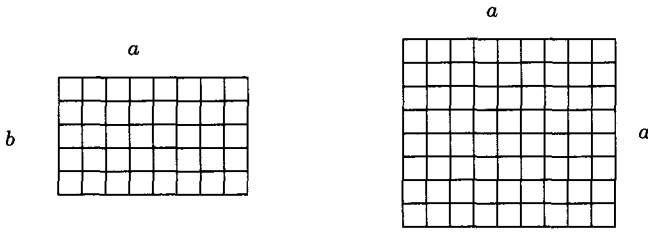


Figure 1.2. Area of a rectangle.

with ratio m/n , then the pairs $BC-EF$ and $AC-DF$, are commensurable both with ratio m/n .

Proof. Since all angles are equal, possibly after a reflection, translation or rotation, we can assume that the two triangles have a common angle $\angle DEF = \angle ABC$, and that the lines AC and DF are parallel, see Figure 1.1; we can moreover assume that A and C are interior points of the segments DC and EF . The commensurability assumption yields a segment e with the property that AB is a multiple of e with a factor m and DE is a multiple of e with a factor n . This way AB and DE are subdivided respectively into n and m pieces equal to e . If we draw the parallel lines to AC through the point of subdivision of DE , we obtain a subdivision of BC and EF respectively into m and n equal pieces. Such a quantity, which is common to BC and EC , is the common measure we were looking for.

Similarly, we can show that AC and DF are commensurable. \square

1.2 Theorem (Area of a Rectangle). Let R be a rectangle with sides a and b , which are commensurable with ratio m/n with respect to a segment e . Then R is commensurable to the square Q of side a with ratio m/n .

Proof. In fact we have $a = ne$ and $b = me$ and, compare Figure 1.2, $R = nme \times e$ and $Q = n^2 e \times e$. \square

1.3 Remark. It is the commensurability of the sides of a rectangle that allows us to measure the area in an elementary way.

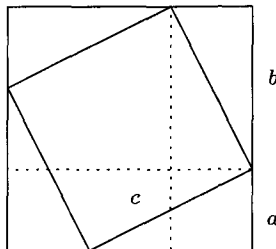


Figure 1.3. Pythagorean theorem: $c^2 + 2ab = (a + b)^2$.

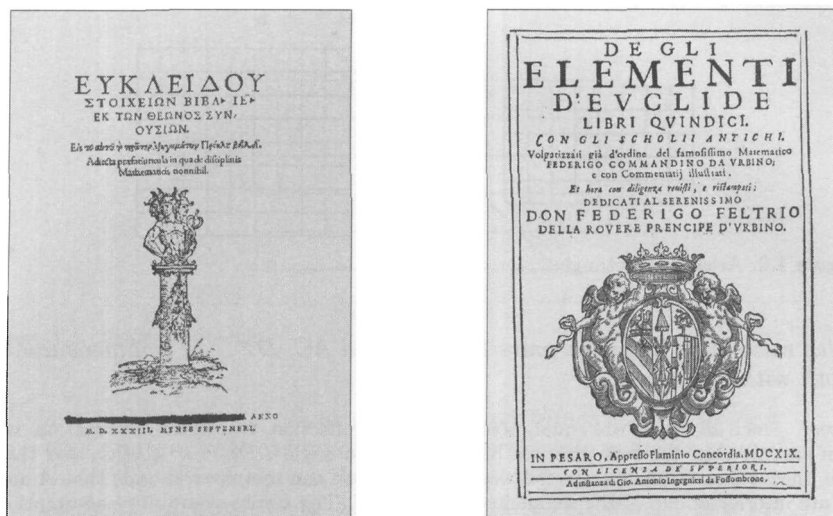


Figure 1.4. Frontispieces of the first printed Greek and Latin editions of the *Elements* by Euclid of Alexandria (325BC–265BC).

1.4 ¶. Show the geometric form of the Pythagorean theorem. That is, show with a straight edge and compass construction that, in a right triangle, we can decompose the square on the sides into parts which fit exactly into the square of the hypotenuse.

1.5 Theorem (Pythagorean theorem). Suppose that the sides and the hypotenuse of a right triangle are commensurable to a segment e with ratios respectively m/n , p/q and r/s ; then

$$\frac{m^2}{n^2} + \frac{p^2}{q^2} = \frac{r^2}{s^2}.$$

Proof. The squares of the sides are commensurable to the square of side e with ratio respectively m^2/n^2 , p^2/q^2 and r^2/s^2 . The claim then follows from the geometric version of the Pythagorean theorem in Exercise 1.4, if we take into account that submultiples of a given quantity are commensurable. \square

b. Never-ending processes

The Pythagorean assumption that all pairs of homogeneous quantities are commensurable was probably supported by proofs such as the ones we have seen in the previous paragraph. The discovery of *incommensurable* pairs of segments, such as the side and the diagonal of a square (see Proposition 1.9 of [GM1]), and its disclosure by Hyppasus, a member of the Pythagorean school, produced a deep crisis in the numerical foundations of geometry and on some of the dominant Greek culture, so much so that it cost the life of Hyppasus himself: according to the tradition, Hippasus was thrown overboard by the Pythagoreans. Obviously, it was not only a mathematical foundation that was failing, but a whole conception of the world, a conception meant to justify social relationships and cultural superiority.

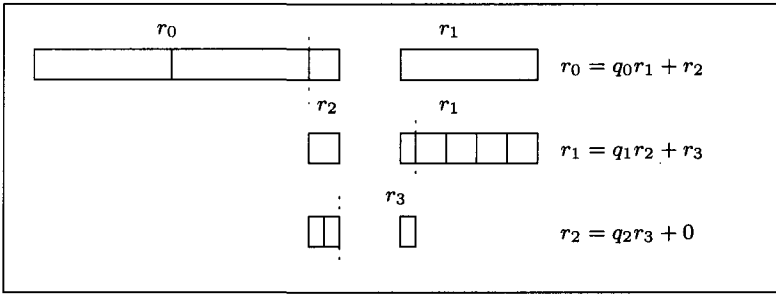


Figure 1.5. Euclid's algorithm.

The procedure for determining a possible unit common to two *magnitudes*, as line segments, angles, areas, can be regarded as the geometric equivalent of *Euclid's algorithm* (see, for example, 8.25). In the case of two line segments r_0, r_1 , we consider the shortest one, r_1 , and we cover r_0 with copies of r_1 . If we succeed in covering r_0 perfectly, r_1 is a common unit as r_0 is a multiple of r_1 . Otherwise, we consider the part r_2 which remains from r_0 after covering it with copies of r_1 , and restart the process using r_2 as the shortest segment between, this time, r_1 and r_2 (see, for example, Figure 1.5). For the Pythagoreans this process would always stop after a finite number of steps.³ In fact stopping after a finite number of steps is exactly equivalent to commensurability. However, the procedure will never stop in the case of the diagonal and the side of a square, as we have seen, or of the diagonal and the side of a pentagon (see, for example, Figures 1.6 and 1.7).

The existence of incommensurable pairs made it necessary to face processes that were treacherous as they did not stop after a finite number of steps, and to give up the idea of controlling continuous geometrical entities by rational numbers or finite processes.

These reasons probably led Eudoxus of Cnidus (408BC–355BC) to introduce the notion of *magnitude* as opposed to numbers and develop a theory of *comparison of magnitudes*: the *theory of proportions* which is presented in Book V of Euclid's *Elements*. This, together with the *method of exhaustion*, due also to Eudoxus, is among the greatest achievements of Euclidean geometry. The method of exhaustion is presented in Book XII of Euclid's *Elements* and finds its splendour with Archimedes of Syracuse (287BC–212BC) and, later, with mathematicians in the Renaissance, as for example Francesco Maurolico (1494–1575).

Eudoxus's idea is to consider *equal ratios of pairs of magnitudes* without any reference to numbers. Nonending processes this way disappear and we capture their essence via the method of exhaustion. Using modern notation and numbers in a nonessential way, we can state

³ In this context one can think of Zenon's paradoxes (about 495BC).

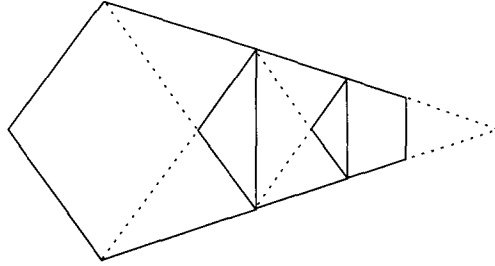


Figure 1.6. Denote by P_n , $n = 1, 2, \dots$, the n -th pentagon from the left to the right and by a_n and d_n respectively the lengths of its side and diagonal. Then we have $d_{n+1} = a_n$ and $a_n = d_{n+1} = a_{n+1} + a_{n+2}$. The figure shows that the process of construction of pentagons never ends: a_1 and d_1 are therefore incommensurable (see, for example, Chapter 3).

1.6 Definition (Exhaustion principle). *The magnitudes a and b are in the same ratio of the magnitudes A and B if, given arbitrarily two positive numbers m and n , we have*

$$\begin{aligned} ma < nb & \quad \text{if and only if} \quad mA < nB, \\ ma > nb & \quad \text{if and only if} \quad mA > nB. \end{aligned}$$

Of course the previous criterion requires “infinitely many comparisons of capability,” but it provides a firm foundation of geometry; for instance, the proofs of Theorems 1.1, 1.2 and 1.5 extend easily to cover “irrational ratios.” Though historically not correct, we can think of the exhaustion method as a method for approximating irrational numbers by rationals.⁴

c. Back to numbers

In Medieval times the centrality of the numbers came up again because of the new trading. The *algebra* brought from the Arab world by Leonardo Pisano (1170–1250), called Fibonacci, took a relevant role in the new commercial companies: any good is homogeneous to any other good, money is the unit of measure to which every quantity has to be referred. New problems, which require numerical solutions, arose and the continuity problem came again as the problem of finding square or cubic roots. Scientists recognized the irrational character of those numbers, but, unlike the Greeks, they learned how to live with them.

They avoided asking themselves about the nature of these new numbers and were satisfied with approximations whenever irrationals appeared as solutions of problems.

⁴ Notice that, if the ratio of magnitudes are numbers, the exhaustion principle is a process that leads to the equality $\frac{a}{b} = \frac{A}{B}$ and amounts to showing that

$$\left| \frac{a}{b} - \frac{A}{B} \right| < \frac{1}{n} \quad \forall n \in \mathbb{N}, n \geq 1.$$

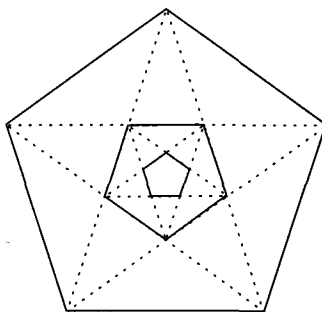


Figure 1.7. The side and the diagonal of a pentagon are incommensurable.

For centuries irrational numbers were used. Meanwhile other numbers appeared. In the fifteenth century the Italian mathematicians Niccolò Fontana (1500–1557), called Tartaglia, Girolamo Cardano (1501–1576) and Rafael Bombelli (1526–1573) used even imaginary numbers in solving algebraic equations of third and fourth degree, and François Viète (1540–1603) introduced literal calculus. The bursting impact of the infinitesimal calculus led to include even “infinity” and “infinitesimal” among numbers. Of course the development of mathematics, especially in the sixteenth and seventeenth centuries did not go without criticism, but in some sense, D’Alembert’s attitude *allez de l’avant: la foi vous viendra* mattered more.

At the beginning of the eighteenth century, Augustin-Louis Cauchy (1789–1857) tried to give solid bases to *infinitesimal calculus*, founding it on the *theory of limits*, that he rigourously developed in two celebrated treatises: the *Cours d’Analyse* and the *Resumé des leçons sur le calcul infinitésimal*, respectively in 1821 and 1823. However in this process of revision he found a series of difficulties that could be overcome, as we have seen in [GM1], only after a rigorous settlement of the system of real numbers.

It was only fifty years later in 1872 that Georg Cantor (1845–1918) and Richard Dedekind (1831–1916) formulated the *axiom of continuity* (see, for example, Section 1.2) and built a *model* of real numbers in the celebrated works *Über die Ausdehnung eines Satzes aus der Theorie der trigonometrischer Reihen* and *Stetigkeit und irrationale Zahlen*.

The system of numbers needed should be a *minimal extension* of the rationals, so that each number could be approximated by rationals. Also, in such a system, we should be able to compare, sum and multiply as one does with the geometrical continuum, but without any reference to it.

The crucial property singled out by Dedekind to capture the intuition of the continuity of the line was that, in every division of the line into two classes of points such that every point in one class is to be to the left of each point in the second, there is *one and only one point* that produces the division. He carried this idea over to the existence of the supremum of every nonempty subset that is bounded from above.

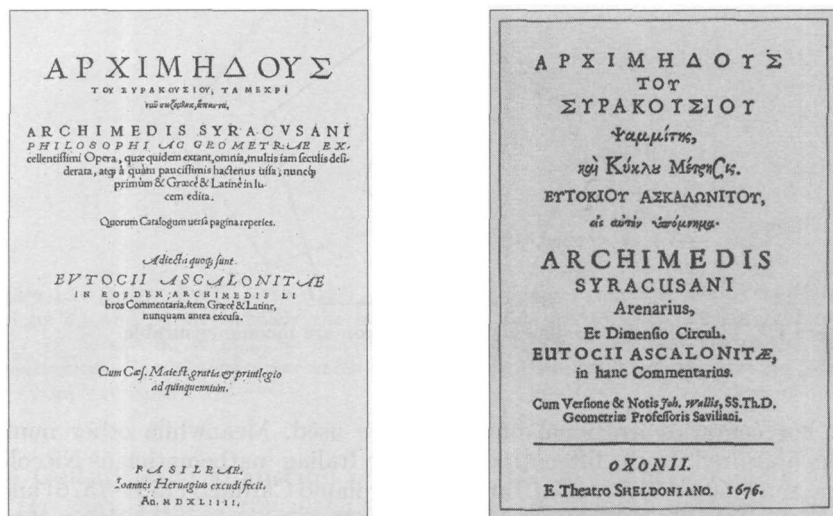


Figure 1.8. Frontispieces of Johannes Herwangen *Editio Princeps* in Greek and Latin of the works of Archimedes of Syracuse (287BC–212BC), Basel 1594, and of one of the Oxford editions, Oxford 1696.

d. An axiomatic or a constructive approach?

The clarification of the mystery of the continuity of the real line due to Georg Cantor (1845–1918) and Richard Dedekind (1831–1916) turned out to be simple and consistent with the way mathematicians had dealt with real numbers in those years. However, the question of the existence of such a system of numbers still held. Actually, immediately after Dedekind's works, other models of real numbers appeared. They were built starting from the rationals as, for instance, the one due to Karl Weierstrass (1815–1897). The idea that became dominant from then on was the following. Starting from the rationals one adds new numbers such that, if one chooses a reference on the line, they will occupy the holes left out by the rationals. Then, by using the possibility of approximating the new numbers with rational numbers, the operations already defined on the rationals are extended to the former as well.

The constructive approach brings back the existence of the system of real numbers, i.e., the consistency of such a system, to the consistency of the rationals and therefore to the one of natural numbers, in a process of “arithmetization of mathematics” typical of the so-called Berlin school around the middle of the nineteenth century, well expressed by the famous words of Leopold Kronecker (1823–1891) :

Natural numbers are the work of God, all else is the work of man.

Actually this process is not at all simple and requires a theory of sets, that is quite abstract and complex. Furthermore, it turns out that within such a

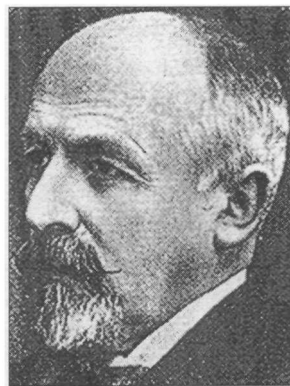
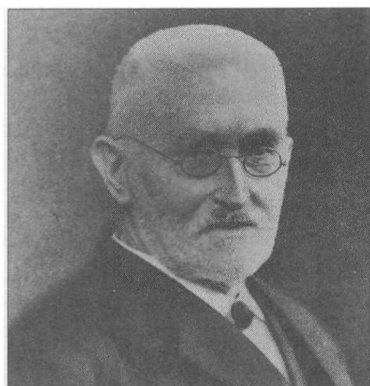


Figure 1.9. Richard Dedekind (1831–1916) and Georg Cantor (1845–1918).

theory, one cannot establish whether these systems are consistent (Gödel's theorem, Kurt Gödel (1906–1978)); in other words, one cannot establish whether the assumption that a system enjoys a number of properties will lead or not to unpleasant surprises: this is the question of the *foundations of mathematics* (see, for example, Section 3.3.2).

We chose in [GM1], and we will insist in our choice in Section 1.2, an axiomatic approach to real numbers: we take for granted that there is a system of numbers that enjoys the properties it is expected to have, and within this system we shall find the subsets of rational and natural numbers.

1.2 The Axiomatic Approach to Real Numbers

In this section we discuss the axioms of the system of real numbers and some of their consequences. For the sake of convenience we deal with *algebraic* and *order* properties in Section 1.2.1, and with the *continuity property* in Section 1.2.2.

1.2.1 Algebraic and order properties

The algebraic properties of real numbers are conveniently subsumed in a minimal number of *axioms* that give the rules of computation. Those are enough to allow us to *derive* the usual rules of computation.

a. Axioms for addition

An operation of sum is defined in the system of real numbers \mathbb{R} : it associates to each pair of numbers x and y their *sum* denoted by $x + y$. In other words a function is defined, the sum, $+: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, $(x, y) \rightarrow x + y$, $x, y \in \mathbb{R}$. We assume that

- (A₁) *Addition is associative*: $(x + y) + z = x + (y + z)$ for all $x, y, z \in \mathbb{R}$.
- (A₂) *Existence of zero*: \mathbb{R} has an element, denoted by 0, such that $0 + x = x + 0 = x$ for every $x \in \mathbb{R}$.
- (A₃) *Existence of the opposite*: to every $x \in \mathbb{R}$ corresponds an element $y \in \mathbb{R}$ such that $x + y = y + x = 0$.
- (A₄) *Addition is commutative*: $x + y = y + x$ for all $x, y \in \mathbb{R}$.

From (A₁), ..., (A₄) we infer, for example,

- (i) The number 0 in (A₂), called *zero* or *neutral element* for the addition, is *unique*. In fact, for another 0' we infer $0 = 0 + 0' = 0'$ by applying (A₂) to 0, A₄ and again (A₂) to 0'.
- (ii) The opposite y in (A₃) to x is *unique*. In fact, if for $y, z \in \mathbb{R}$ we had $x + z = x + y = 0$, then $z = z + 0 = z + (x + y) = (z + x) + y = 0 + y = y$ by applying (A₂), (A₃), (A₁), (A₂) and again (A₁). The opposite of x is usually denoted by $-x$ and one writes $x - y$ instead of $x + (-y)$. The new operation $(x, y) \rightarrow x - y$ is then called *subtraction*.

Whenever in a set X an operation with the properties (A₁), ..., (A₄) is defined, we say that X is a *commutative group*. In this case, (i) and (iii) above read: in a commutative group there is a unique neutral element, and every x has a unique opposite. Axioms (A) for the reals can therefore be summarized by saying that \mathbb{R} is a *commutative group with respect to addition*.

b. Axioms for multiplication

A second operation, called *multiplication*, $(x, y) \rightarrow xy$, $\forall x, y \in \mathbb{R}$, is assumed on \mathbb{R} . It satisfies the following axioms:

- (M₁) *Multiplication is associative*: $(xy)z = x(yz)$ for all $x, y, z \in \mathbb{R}$.
- (M₂) *Existence of identity*: \mathbb{R} contains an element, denoted by 1, such that $1 \neq 0$ and $1x = x1 = x$ for every $x \in \mathbb{R}$.
- (M₃) *Existence of the reciprocal*: to each $x \in \mathbb{R}$, $x \neq 0$, corresponds an element $w \in \mathbb{R}$ such that $wx = xw = 1$.
- (M₄) *Multiplication is commutative*: $xy = yx$ for all $x, y \in \mathbb{R}$.

Similarly to addition one easily proves that the identity is unique and the reciprocal of each element is unique. Usually one denotes by x^{-1} , $1/x$ or by $\frac{1}{x}$ the reciprocal of $x \neq 0$. We emphasize that 0^{-1} or $1/0$ is not defined and it is meaningless. It is easily seen that $\mathbb{R} \setminus \{0\}$ is a commutative group with respect to multiplication.

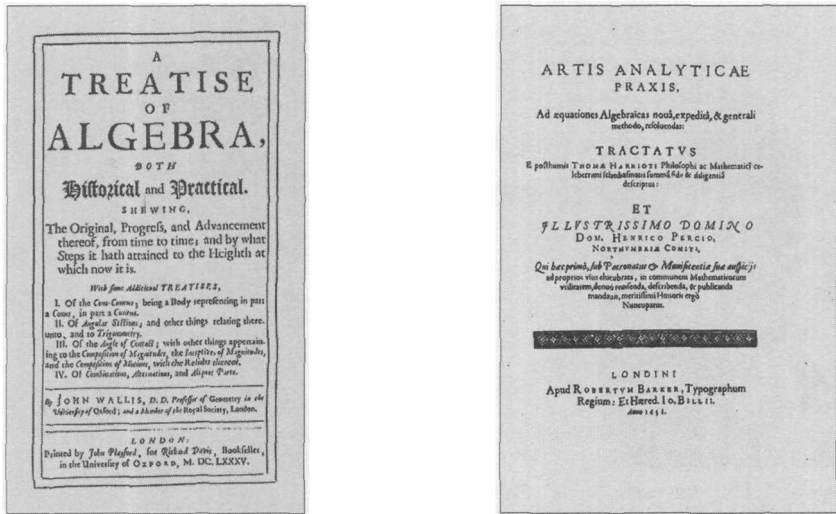


Figure 1.10. Frontispieces of *A treatise of Algebra* by John Wallis (1616–1703) and of *Artis analyticae praxis* by Thomas Herriot (1560–1621) where probably the symbols $<$ and $>$ first appear.

1.7 ¶. Show that rotations of the plane around a given point form a commutative group, the operation of sum of two rotations, respectively of angles x and y , being defined as the rotation of angle $x + y$. Since we can clearly identify rotations of the plane with the unit circle in \mathbb{R}^2 , we can say in a fancy way that the circle has the structure of a commutative group.

1.8 ¶. Show that rotations of the space around a given point form a group which is however not commutative, that is, rules (A_1) , (A_2) , (A_3) , but not (A_4) hold. Again in a fancy way we can say that the unit sphere in \mathbb{R}^3 has the structure of a noncommutative group.

c. The distributive law

The next axiom defines the relationship between the operations of sum and multiplication.

$$(AM) \quad x(y + z) = xy + xz \text{ holds for all } x, y, z \in \mathbb{R}.$$

All algebraic rules of computation follow from the axioms (A) , (M) and (AM) .

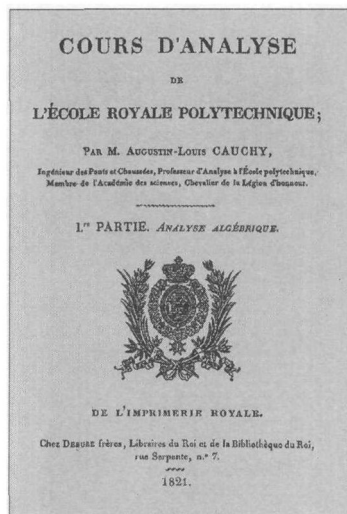
1.9 ¶. Show that

- (i) $0 \cdot x = 0$,
- (ii) $(-x)(-y) = xy$,
- (iii) $(-x)y = -(xy)$,
- (iv) $(x - y)z = xz - yz$,
- (v) $xy = 0$ if and only if either $x = 0$ or $y = 0$.

[Hint: As an example let us prove (i). We have $0 \cdot x + x = 0x + 1x$ [by (M_2)] = $(0 + 1)x$ [by (AM) and (M_4)] = $1x$ [by (A_2)] = x [by (M_2)]. Summing to both sides $-x$ we then infer $0x = 0x + (x + (-x)) = (0x + x) + (-x)$ by $(A_1) = x + (-x) = 0$ by (A_3) .]



Figure 1.11. Augustin-Louis Cauchy (1789–1857) and the frontispiece of his *Cours d'Analyse*, 1821.



If $x, y \in \mathbb{R}$, $x \neq 0$ the *quotient* of y by x is defined as

$$\frac{y}{x} := y \frac{1}{x} = yx^{-1}.$$

We also write y/x for $\frac{y}{x}$, $x \neq 0$. The ordinary rules of computation of *fractions*, as for instance

$$\frac{a}{b} \frac{c}{d} = \frac{ac}{bd}, \quad \frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}$$

follow easily from the axioms for multiplication. We repeat: *dividing by zero is not allowed*.

d. Order

We can identify in \mathbb{R} a subset P , called the subset of *positive numbers*, by means of the following two axioms:

- (O_1) If x, y are positive numbers, $x, y \in P$, then $x + y$ and $xy \in P$.
- (O_2) For each $x \in \mathbb{R}$ only one of the following three alternatives holds:
 $x \in P$, $x = 0$ or $-x \in P$.

(O_1) and (O_2) imply that 1 is positive. In fact, since $1 \neq 0$, either 1 or -1 is positive and, as $1 = 1^2 = (-1)^2$ we conclude that 1 is positive. A nonzero number which is nonpositive, is called *negative*. We write $x > 0$ to say that x is positive, while $x > y$ or $y < x$ mean that $x - y$ is positive. Consequently $x < 0$ means that x is negative, and x negative, $x < 0$, is equivalent to $-x$ is positive, $-x > 0$. One can show that if x, y are negative,

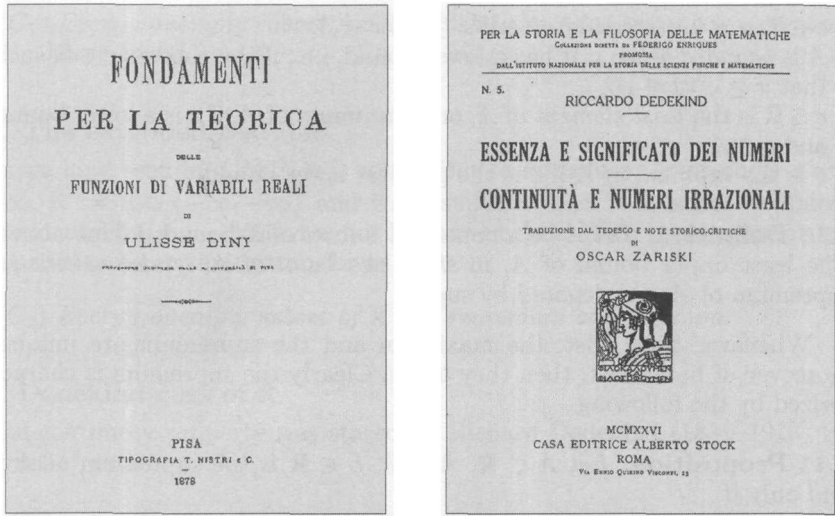


Figure 1.12. Frontispieces of the *Fondamenti per la teorica delle funzioni di variabili reali* by Ulisse Dini (1845–1918) and of the Italian translation by Oscar Zariski (1899–1986) of *Stetigkeit und irrationale Zahlen* by Richard Dedekind (1831–1916).

then xy is positive. In fact $xy = (-x)(-y)$ and $-x$ and $-y$ are positive. In particular *the square of a nonzero real number is positive*.

From the previous axioms it is not difficult to infer the usual rules to deal with inequalities:

- (i) if $x < y$ and $y < z$, then $x < z$,
- (ii) if $x < y$ and $z > 0$, then $xz < yz$,
- (iii) if $x < y$ and $z \in \mathbb{R}$, then $x + z < y + z$,
- (iv) if $x < y$ and $x > 0$, then $\frac{1}{y} < \frac{1}{x}$,
- (v) if $x < y$ and $z < 0$, then $xz > yz$.

Finally, since 1 is positive, also $2 := 1 + 1$, $3 := 1 + 1 + 1$, $1 + 1 + \cdots + 1$, and so on are positive.

1.2.2 Continuity property

a. Supremum

Let A be a nonempty subset of \mathbb{R} . We recall (see, for example, Section 1.1 of [GM1])

- $c \in \mathbb{R}$ is an *upper bound* of A if $A \subset]-\infty, c]$, i. e., if $x \leq c \forall x \in A$,
- A is *bounded above* if it has an upper bound, i.e., if there exists $c \in \mathbb{R}$ such that $x \leq c \forall x \in A$,
- c is the *greatest element* of A , or a *maximum* of A , if c is an upper bound of A and $c \in A$,

- $c \in \mathbb{R}$ is a *lower bound* of A if $x \geq c \ \forall x \in A$,
- A is *bounded below* if it has a lower bound, i.e., if there exists $c \in \mathbb{R}$ such that $x \geq c \ \forall x \in A$,
- $c \in \mathbb{R}$ is the *least element* of A , or a *minimum* of A , if c is a lower bound and $c \in A$,
- c is the *infimum* of A if c is the greatest lower bound.

1.10 Definition. Let A be a nonempty subset of \mathbb{R} bounded from above. The least upper bound of A , in short the l.u.b. of A , is also called the supremum of A and denoted by $\sup A$.

Whenever they exist, the maximum and the supremum are unique, moreover, if both exist, then they agree. Clearly the supremum is characterized by the following

1.11 Proposition. Let $A \subset \mathbb{R}$, $A \neq \emptyset$. $L \in \mathbb{R}$ is the supremum of A if and only if

- (i) L is an upper bound of A , i.e., $x \leq L \ \forall x \in A$,
- (ii) $\forall \epsilon > 0 \ L - \epsilon$ is not an upper bound of A , i.e., $\forall \epsilon > 0 \ \exists x \in A$ such that $x > L - \epsilon$.

The *axiom of continuity* of the reals is then (see, for example, Section 1.1 of [GM1])

(C) Every nonempty subset of \mathbb{R} that is bounded above has a least upper bound.

1.12 Remark. If c is an upper bound of A , then every number larger than c is again an upper bound of A . We are tempted to say that the upper bounds of A form a half-line, but to identify it we need the left extremal point! A geometric way of visualizing the axiom of continuity is exactly saying that if $A \subset \mathbb{R}$ is nonempty and bounded above, then all upper bounds of A are given by the numbers in $[\sup A, +\infty[$.

1.13 Example. If $A =]-\infty, c]$, c is the maximum of A . All upper bounds of A are given by the numbers in the closed half-line $[c, +\infty[$, and $c = \sup A$.

If $A =]-\infty, c[$, A has no maximum, all upper bounds of A are again the numbers in the closed half-line $[c, +\infty[$, and c is the supremum of A . In both cases the set of upper bounds is given by the closed half-line $[c, +\infty[$, and the l.u.b. is c , therefore it exists.

Similarly we have

1.14 Proposition. Let $A \subset \mathbb{R}$, $A \neq \emptyset$. $L \in \mathbb{R}$ is the infimum of A if and only if

- (i) L is a lower bound of A , i.e., $L \leq x \ \forall x \in A$,
- (ii) $\forall \epsilon > 0 \ L + \epsilon$ is not a lower bound of A , i.e., $\forall \epsilon > 0 \ \exists x \in A$ such that $x < L + \epsilon$.

Equivalently, the axiom of continuity can be restated as

(C₁) Every nonempty subset $A \subset \mathbb{R}$ which is bounded below has a greatest lower bound.

b. The extended real line

As we have seen in [GM1], it is convenient to introduce the symbols $+\infty$, $-\infty$, $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty, -\infty\}$ and write $\sup A = +\infty$ if A is not bounded above, and $\inf A = -\infty$ if A is not bounded below. With these agreements the axiom of continuity transforms into

(C₂) Every nonempty subset of \mathbb{R} has supremum and infimum.

c. Dedekind cuts of \mathbb{R}

The continuity property was stated by Richard Dedekind (1831–1916) in terms of *cuts*.

1.15 Definition. Let X be a set in which the axioms (A), (M), (AM) and (O) hold. A cut (A, B) of X is a subdivision of X in nonempty subsets A and B such that $A \cup B = X$, $A \cap B = \emptyset$ and

$$\forall a \in A \text{ and } \forall b \in B \text{ we have } a < b.$$

If (A, B) is a cut of X , we say that $x \in X$ corresponds to (A, B) or that it brings about this cut if $a \leq x \leq b \forall a \in A, \forall b \in B$.

Clearly the element that brings about a cut is unique, if it exists, and belongs either to A or to B .

1.16 Theorem. Let X satisfy the axioms (A), (M), (AM) and (O). The following

- (i) the axiom of continuity (C) holds in X ,
- (ii) to every cut of X corresponds an element of X

are equivalent.

Following Dedekind, we can then state the axiom of continuity also as

(C₃) To every cut of X corresponds an element of X .

Proof of Theorem 1.16. (i) \Rightarrow (ii). Let (A, B) be a cut in X . Clearly A is bounded above. Set $x_0 = \sup A$. We show that x_0 brings about the cut (A, B) . Since x_0 is an upper bound, we have $a \leq x_0$ for all $a \in A$. Since x_0 is the least upper bound of A , $x_0 \leq b$ for all $b \in B$.

(ii) \Rightarrow (i). Let E be a nonempty subset of X that is bounded above. Denote by $M(E)$ the set of upper bounds of E . Clearly $A := \mathbb{R} \setminus M(E)$ and $B := M(E)$ form a cut (A, B) of X . Denote by x_0 the element of X corresponding to (A, B) . We then show that x_0 is an upper bound of E . Otherwise, there is an $x \in E$ such that $x > x_0$ and choosing $x_1 := (x_0 + x)/2$, we get $x_1 \in M(E)$ since $x_1 > x_0$ and $x_1 \notin M(E)$ since $x_1 < x \in E$, a contradiction. Since moreover $x_0 \leq b \forall b \in M(E)$, x_0 is the l.u.b. of E . \square

The axiom of continuity is not valid in \mathbb{Q} . For example, if $A := \{x \in \mathbb{Q} \mid 0 \leq x^2 < 2\}$, A is nonempty, bounded above, and $\sup A = \sqrt{2} \in \mathbb{R}$; but, being that $\sqrt{2}$ is not in \mathbb{Q} , A has no supremum in \mathbb{Q} .

1.17 ¶. The existence of the n -th root of a nonnegative real number is a consequence of the continuity of the function x^n , $x > 0$ (see, for example, 2.46 of [GM1]). Relying on the axiom of continuity, prove that every nonnegative real number has an n -th root.

1.2.3 Uniqueness of reals

We have already hinted at the fact that it cannot be decided whether the system of reals is consistent or not. Another important question is the uniqueness of such a system. This is not clear a priori; both the rationals and the reals satisfy the algebraic and order axioms, but $\mathbb{Q} \neq \mathbb{R}$. Fortunately two numerical systems S and T which satisfy the algebraic, order *and* continuity axioms are undistinguishable. Let us be more precise on this point.

A set S with two operations, called addition and multiplication, which satisfy the axioms (A) , (M) , (AM) , and (0) , is called an *ordered field*. For example \mathbb{R} and \mathbb{Q} are ordered fields. An ordered field is said to be *complete* if the axiom of continuity holds.

An *algebraic isomorphism* $f : S \rightarrow S'$ is a bijective correspondence between S and S' that is compatible with the operations of addition and multiplication on S and S' , that is, such that

$$f(x +_S y) = f(x) +_{S'} f(y), \quad f(x \cdot_S y) = f(x) \cdot_{S'} f(y)$$

for all $x, y \in S$. We say that the isomorphism is *order preserving* when $f(x)$ is positive in S' if and only if x is positive on S , i.e.,

$$x <_S y \text{ if and only if } f(x) <_{S'} f(y) \quad \forall x, y \in S.$$

Finally, we say that the ordered fields S and S' are *isomorphic* if there is an algebraic isomorphism $f : S \rightarrow S'$ that preserves the order. We then have

1.18 Theorem. *Every complete ordered field S is isomorphic to \mathbb{R} , consequently all complete ordered fields are isomorphic.*

Proof. Since S is complete, any of its bounded subsets has supremum in S . Also, if $f : S \rightarrow S'$ is an isomorphism between two complete ordered fields S and S' , then we have $f(\sup E) = \sup f(E)$.

We can now easily construct a bijection $f : \mathbb{N} \rightarrow S$ between \mathbb{N} and a subset of S which preserves operations and order;⁵ we can also extend this bijection to a bijection f of \mathbb{R} onto a subset $S' \subset S$. To conclude, it suffices to prove that $f(\mathbb{R}) = S$. Suppose $f(\mathbb{R}) \neq S$, i.e., that there is $\bar{x} \in S \setminus S'$ and let $\mathbb{Q}' := f(\mathbb{Q})$, and consider the sets

$$\{x \in \mathbb{Q}' \mid x < \bar{x}\} \quad \text{and} \quad \{x \in \mathbb{Q}' \mid x > \bar{x}\}.$$

⁵ Compare the next section concerning the subset of naturals in \mathbb{R} .

They are clearly nonempty, otherwise \mathbb{Q}' would be bounded below or above. If

$$\ell' := \sup\{x \in \mathbb{Q}' \mid x < \bar{x}\}, \quad \text{and} \quad L' := \inf\{x \in \mathbb{Q}' \mid x > \bar{x}\},$$

we have $\ell', L' \in S'$ and

$$L' < \bar{x} < \ell',$$

otherwise $\bar{x} \in S'$. Therefore we conclude that there is $p \in \mathbb{Q}'$ such that $L' < p < \ell'$; this is a contradiction, as it would imply that there are no rationals between ℓ and L , where $\ell' := f(\ell)$, $L' := f(L)$. \square

1.3 Natural Numbers

In this section we identify the subset of \mathbb{R} of natural numbers.

a. Natural numbers and the principle of induction

We commonly say that natural numbers are the numbers 0, 1, 2, 3, and so on, actually meaning that there is a never-ending rule producing *all* natural numbers which is

- (i) 0 is a natural number,
- (ii) if x is a natural number, then adding 1 produces the next natural number, the “successor” $x + 1$ of x .

Even more, we intend that this rule generates *only* natural numbers.

To be more precise, let us state first

1.19 Definition. A subset $A \subset \mathbb{R}$ is said to be inductive if

- (i) $0 \in A$,
- (ii) if $x \in A$, then $x + 1 \in A$.

The entire \mathbb{R} , the half-lines $[-1, \infty[$ and $[0, \infty[$, the subset

$$\left\{0, 1/2, 1, 3/2, 2, 5/2, 3, \dots\right\}$$

are examples of inductive subsets of \mathbb{R} . The naive way to describe the naturals suggests that

- (i) the set of natural numbers is inductive,
- (ii) no proper subset of the naturals is inductive,
- (iii) the subset of naturals is the smallest inductive subset of \mathbb{R} .

For these reasons we set

1.20 Definition. \mathbb{N} is the smallest inductive subset of \mathbb{R} .

A trivial consequence is the following.

1.21 Proposition (Induction principle). *If $A \subset \mathbb{N}$ is inductive, then $A = \mathbb{N}$.*

The Definition 1.20 can be justified by naive set theory. In fact we can define \mathbb{N} as the intersection of *all* inductive subsets of \mathbb{R} ,

$$\mathbb{N} := \bigcap \{A \subset \mathbb{R} \mid A \text{ is inductive}\}.$$

This way the existence of \mathbb{N} leads to set theory. Then we need to show that \mathbb{N} is inductive itself. Consequently \mathbb{N} exists and is the smallest inductive subset of \mathbb{R} .

1.22 ¶¶. Show that $\mathbb{N} := \bigcap \{A \subset \mathbb{R} \mid A \text{ inductive}\}$ is inductive.

To be consistent, it remains to show that the operations of \mathbb{R} , when restricted to \mathbb{N} , yield the usual operations on \mathbb{N} . This is summarized in the following

1.23 Proposition. *We have:*

- (i) *if $n \in \mathbb{N}$, then $n + 1 \in \mathbb{N}$,*
- (ii) *if $n, m \in \mathbb{N}$, then $n + m$ and $nm \in \mathbb{N}$,*
- (iii) *if $n \in \mathbb{N}$ and $n > 0$, then $n - 1 \in \mathbb{N}$,*
- (iv) *if $n, m \in \mathbb{N}$ and $|n - m| < 1$, then $n = m$,*
- (v) *every nonempty subset $A \subset \mathbb{N}$ has a minimum,*
- (vi) *a subset $A \subset \mathbb{N}$ is bounded if and only if it has a maximum.*

Proof. Notice that in principle $n + 1$, $n + m$, nm , are real numbers.

(i) It is trivial, since \mathbb{N} is inductive.

(ii) For $n \in \mathbb{N}$ set $A_n := \{m \in \mathbb{N} \mid n + m \in \mathbb{N}\}$. It is easily seen that A_n is an inductive subset of \mathbb{R} . Thus by the induction principle $A_n = \mathbb{N}$, that is $n + m \in \mathbb{N} \forall m$ and fixed n . The claim follows since n is arbitrary. One can argue similarly for the product of naturals.

(iii) The set $A := \{0\} \cup \{n \in \mathbb{N} \mid n - 1 \in \mathbb{N}\}$ is inductive, hence $A = \mathbb{N}$, in particular $n - 1 \in \mathbb{N}$ if $n \neq 0$.

(iv) We claim that the set $A := \{n \in \mathbb{N} \mid \nexists m \in \mathbb{N}, n < m < n + 1\}$ is inductive. In fact, if we found a natural number m with $0 < m < 1$, then $m - 1 < 0$ and (iii) would give $m - 1 \in \mathbb{N}$, i.e., a contradiction. Similarly, we show that if there is no natural number between $n \in \mathbb{N}$ and $n + 1$, then there is no natural number between $n + 1$ and $n + 2$. By the induction principle, $A = \mathbb{N}$.

(v) Let $\ell := \inf A$. If ℓ is not the minimum of A , we can find (by the properties of the infimum) $x, y \in A \subset \mathbb{N}$ with $\ell < y < x < \ell + 1/2$. A contradiction to (iv).

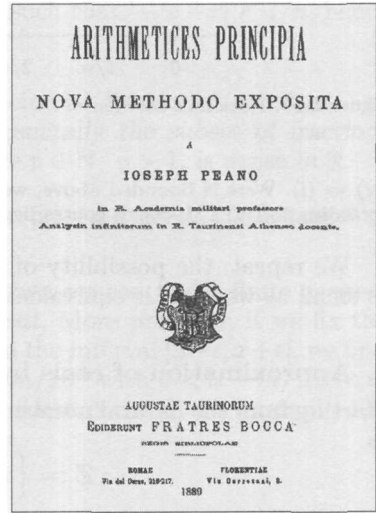
(vi) Let $A \subset \mathbb{N}$ be bounded and let $\ell := \sup A \in \mathbb{R}$. Suppose ℓ is not a supremum, then there are $n, m \in A$ such that $\ell - 1 < n < m < \ell$. Since $A \subset \mathbb{N}$, we reach a contradiction to (iv). \square

1.24 Axiomatic definition of naturals. Natural numbers can also be defined axiomatically independently from the reals. They are a set \mathbb{N} with an application $\sigma : \mathbb{N} \rightarrow \mathbb{N}$ called *successor*, satisfying the following five axioms:

- (i) $0 \in \mathbb{N}$,



Figure 1.13. Giuseppe Peano (1858–1932) and the frontispiece of *Arithmetices Principia*, Torino, 1889.



- (ii) if $a \in \mathbb{N}$, then $\sigma(a) \in \mathbb{N}$,
- (iii) if $a \in \mathbb{N}$ and $a = \sigma(b)$, then $a \neq 0$,
- (iv) σ is injective, i.e., if the successors of a and b are equal, then so are a and b ,
- (v) if $A \subset \mathbb{N}$ is such that $0 \in A$, and $a \in A$ implies $\sigma(a) \in A$, then $A = \mathbb{N}$.

Axioms (i)–(v) were introduced by Giuseppe Peano (1858–1932), who also showed how one can derive from them the entire arithmetic: they are known as *Peano's axiom*. Starting from natural numbers one can build successively the system of signed integers, denoted by \mathbb{Z} , of rationals \mathbb{Q} and of reals \mathbb{R} .

From Proposition 1.23 (vi) we in particular infer

1.25 Theorem (Archimedean property). \mathbb{N} is not bounded above in \mathbb{R} , i.e., given any $M \in \mathbb{R}$ there exists $n \in \mathbb{N}$ such that $n \geq M$.

It is convenient to state the Archimedean property of \mathbb{R} in several equivalent forms.

1.26 Proposition. The following equivalent claims hold:

- (i) if $M > 0$, then there exists $n \in \mathbb{N}$ such that $n > M$,
- (ii) (ARCHIMEDEAN PROPERTY) if $x, y \in \mathbb{R}$ are positive numbers, then there is $n \in \mathbb{N}$ such that $nx > y$,
- (iii) for every $\epsilon > 0$ there exists $\nu \in \mathbb{N}$ such that $1/\nu < \epsilon$,
- (iv) if $x \in \mathbb{R}$ is such that $|x| < \frac{1}{n}$, $\forall n \in \mathbb{N}$, $n \geq 1$, then $x = 0$.

Proof. (i) is true by Theorem 1.25.

(i) \Rightarrow (ii). It suffices to apply (i) with $M := y/x$.

(ii) \Rightarrow (iii). It suffices to apply (ii) with $y := 1$ e $x := \epsilon$.

(iii) \Rightarrow (iv). If $x \neq 0$, we apply (iii) with $\epsilon := |x|$ and find ν such that $1/\nu < |x|$: a contradiction.

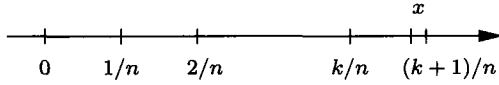


Figure 1.14. Rationals are dense.

(iv) \Rightarrow (i). Were \mathbb{N} bounded above, we would find L such that $L > n \forall n \in \mathbb{N}$, hence, according to (iv) $1/L = 0$: a contradiction. \square

We repeat: the possibility of dividing an interval of \mathbb{R} in subintervals as small as we want is equivalent to the unboundedness of \mathbb{N} in \mathbb{R} .

b. Approximation of reals by rational numbers

Starting from the natural numbers we define the (relative) integer numbers as

$$\mathbb{Z} := \left\{ x \in \mathbb{R} \mid |x| \in \mathbb{N} \right\}$$

and the rational numbers as

$$\mathbb{Q} := \left\{ x \in \mathbb{R} \mid x = \frac{p}{q}, p, q \in \mathbb{N}, q \neq 0 \right\}.$$

1.27 Definition. We say that $A \subset \mathbb{R}$ is dense in \mathbb{R} if for any pair of distinct real numbers $x, y \in \mathbb{R}$, $x < y$, there is $a \in A$ such that $x < a < y$.

Clearly the following claims are equivalent:

- $A \subset \mathbb{R}$ is dense in \mathbb{R} ,
- if $\epsilon > 0$, $x \in \mathbb{R}$, then there is $a \in A$ such that $|x - a| < \epsilon$,
- if $n \in \mathbb{N}$ and $x \in \mathbb{R}$, then there is $a \in A$ such that $|x - a| < 1/n$.

1.28 Theorem. The subset of rationals $\mathbb{Q} \subset \mathbb{R}$ is dense in \mathbb{R} .

Proof. Let $n \in \mathbb{N}$ and let us prove that, if $x > 0$, then there is a rational $r \geq 0$ such that $|r - x| < 1/n$.

- (i) If $0 \leq x < 1/n$, we take $r := 0$ as $|x - 0| = x < 1/n$.
- (ii) If $x \geq 1/n$, we define, compare Figure 1.14,

$$A := \left\{ m \in \mathbb{N}, \left| \frac{m}{n} \leq x \right. \right\}.$$

Since $1 \in A$ as $1/n \leq x$ by the Archimedean property, A is nonempty. Moreover A is bounded above (by nx), hence it has a maximum k . We must have $k \leq nx < (k+1)$, hence

$$\left| x - \frac{k}{n} \right| < \frac{1}{n}.$$

Finally, if $x < 0$ we find $r \geq 0$, $r \in \mathbb{Q}$, such that $|-x - r| < 1/n$, hence $|x - (-r)| = |-x - (r)| < 1/n$. \square

1.29 Theorem. *The subset of decimal fractions, written as $\{m/10^n \mid m \in \mathbb{Z}, n \in \mathbb{N}\}$, and, more generally the subset of fractions of the type $\{m/p^n \mid m \in \mathbb{Z}, n \in \mathbb{N}\}$, where $p \in \mathbb{N}$, $p > 1$, is dense in \mathbb{R} .*

1.30 ¶. Prove Theorem 1.29. [Hint: Compare the proof of Theorem 1.28 and use that, if $n \in \mathbb{N}$ and $p \geq 2$, then $p^n > n$.]

The last theorem says that if $x \in \mathbb{R}$, then we can find a finite decimal expansion $m/10^n$ as close to x as we want. More precisely, if we fix the deviation $\epsilon > 0$ and apply Theorem 1.29 in the interval $|x - \epsilon, x + \epsilon|$, we find an approximate finite decimal expansion $m/10^n$ with $0 < x - m/10^n < \epsilon$. Notice also that not every rational is a finite decimal: for example $1/3$ cannot be expressed as a decimal fraction.

c. Recursive statements

The induction principle has also the following useful formulation.

1.31 Proposition. *Suppose that for every natural number $n \in \mathbb{N}$ we are given a statement $p(n)$.*

- (i) *Suppose that the statement $p(0)$ is known to be true.*
- (ii) *Suppose that for any n , if the statement $p(n)$ happens to be true, then the statement $p(n + 1)$ must also be true.*

Then the statement $p(n)$ must be true for all n .

Proof. Proposition 1.31 is quite convincing: it is equivalent to the induction principle. In fact the assumptions (i) and (ii) just say that the set $A := \{n \in \mathbb{N} \mid p(n) \text{ is true}\}$ is inductive, hence $A = \mathbb{N}$ by the induction principle. \square

Of course we also have

1.32 Proposition. *Suppose that for every n we are given a statement $p(n)$.*

- (i) *Suppose that there is $k \in \mathbb{N}$ such that $p(k)$ is true.*
- (ii) *Suppose that for all n , if the statements $p(k)$, $p(k + 1)$, \dots , $p(n)$ happen to be true, then $p(n + 1)$ must also be true.*

Then the statement $p(n)$ must be also true for all $n \geq k$.

1.33 Example. We show that $2^n \geq n$, $\forall n \geq 0$. Let $p(n)$ be the statement “ $2^n \geq n$.” We have

- (i) $p(0) = “2^0 = 1 \geq 0”$ is true.
- (ii) From “ $2^0 = 1 \geq 0$ ” by adding 1 to both sides we then infer $2^1 = 1 + 1 \geq 1 + 0 = 1$ i.e., $p(1)$ is true and, from $2^n \geq n$ we get $2^{n+1} = 2 \cdot 2^n \geq 2n \geq n + 1$, i.e., $p(n + 1)$ is true.

Proposition 1.31 then yields the estimate $2^n \geq n$ for all n .

1.34 Example (Bernoulli's inequality). We can give a proof of *Bernoulli's inequality* (see, for example, 5.52 of [GM1]) that makes no use of calculus. In fact for $n = 0$ we have $(1 + h)^0 = 1 = 1 + 0 \cdot h$. If now $n \in \mathbb{N}$ and we assume $(1 + h)^n \geq 1 + nh$, then

$$\begin{aligned}(1 + h)^{n+1} &= (1 + h)(1 + h)^n \geq (1 + h)(1 + nh) && [\text{since } h > -1] \\ &= 1 + (n + 1)h + h^2 \geq 1 + (n + 1)h.\end{aligned}$$

1.35 Example (Arithmetic and quadratic mean). Let us show by induction that

$$\left(\frac{1}{n} \sum_{j=1}^n a_j \right)^2 \leq \frac{1}{n} \sum_{j=1}^n a_j^2.$$

For $n = 1$ the claim is trivial. Suppose the claim true for n and let us prove it for $n + 1$. We have

$$\left(\sum_{j=1}^{n+1} a_j \right)^2 = \left(\sum_{j=1}^n a_j \right)^2 + 2a_{n+1} \sum_{j=1}^n a_j + a_{n+1}^2. \quad (1.1)$$

From the inequality $2\alpha\beta \leq \epsilon\alpha^2 + \frac{\beta^2}{\epsilon}$, which holds for all $\alpha, \beta \in \mathbb{R}$ and $\epsilon > 0$, we infer, for $\epsilon := 1/n$, that

$$2a_{n+1} \sum_{j=1}^n a_j \leq na_{n+1}^2 + \frac{1}{n} \left(\sum_{j=1}^n a_j \right)^2. \quad (1.2)$$

Formulas (1.1) (1.2) and the inductive assumption then yield

$$\begin{aligned}\left(\sum_{j=1}^{n+1} a_j \right)^2 &\leq \left(1 + \frac{1}{n} \right) \left(\sum_{j=1}^n a_j \right)^2 + (n + 1)a_{n+1}^2 \leq \left(\frac{n + 1}{n} \right) n \sum_{j=1}^n a_j^2 + (n + 1)a_{n+1}^2 \\ &= (n + 1) \sum_{j=1}^{n+1} a_j^2.\end{aligned}$$

1.36 Example (Sum of the first n naturals). There is a *closed formula* for the sum $S_1(n)$ of the first n naturals.

$$S_1(n) := 1 + 2 + \cdots + n = \sum_{j=1}^n j = \frac{n(n + 1)}{2}. \quad (1.3)$$

This can be proved in several ways.

(i) Writing

$$\begin{aligned}S_1(n) &= 1 + 2 + \cdots + (n - 1) + n, \\ S_1(n) &= n + (n - 1) + \cdots + 2 + 1,\end{aligned}$$

and summing we get,

$$2S_1(n) = (n + 1) + (n + 1) + \cdots + (n + 1) = n(n + 1).$$

(ii) Arranging squares of side 1 as in Figure 1.15, the total area of the shadow squares is

$$S_1(n) = \frac{n^2}{2} + \frac{n}{2} = \frac{n(n + 1)}{2}.$$

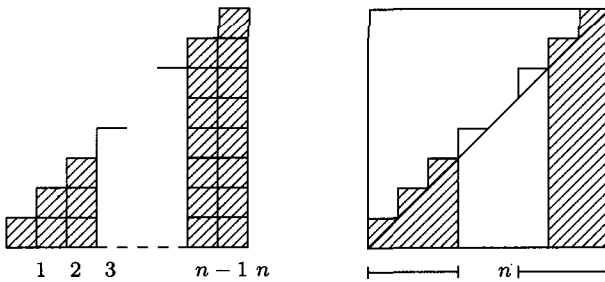


Figure 1.15. $\sum_{j=1}^n j = n(n+1)/2$.

(iii) Using the identity $(a+b)^2 = a^2 + 2ab + b^2$, we can write

$$\begin{aligned}
 1^2 &= & + & + 1 \\
 2^2 &= & 1^2 + & 2 \cdot 1 + 1 \\
 3^2 &= & 2^2 + & 2 \cdot 2 + 1 \\
 &\dots & & \\
 n^2 &= & (n-1)^2 + & 2 \cdot (n-1) + 1 \\
 (n+1)^2 &= & n^2 + & 2 \cdot n + 1
 \end{aligned}$$

and, summing,

$$1^2 + 2^2 + \dots + (n+1)^2 = 1^2 + 2^2 + \dots + n^2 + 2S_1(n) + (n+1),$$

that is, $2S_1(n) = (n+1)^2 - (n+1) = n(n+1)$.

(iv) By induction: the sequence $x_n := n(n+1)/2$, $n \geq 1$, satisfies the recursion

$$\begin{cases} x_1 = 1, \\ x_{n+1} = x_n + (n+1), \quad \forall n \geq 1, \end{cases}$$

which defines $S_1(n)$.

1.37 Example. The sum of the first n odd naturals is

$$\sum_{j=1}^n (2j-1) = 2 \sum_{j=1}^n j - n = n(n+1) - n = n^2,$$

see Figure 1.16.

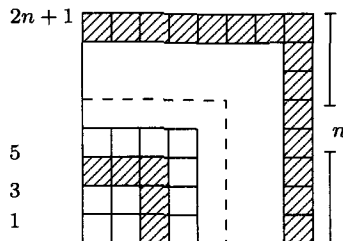


Figure 1.16. $\sum_{j=1}^n (2j-1) = n^2$.

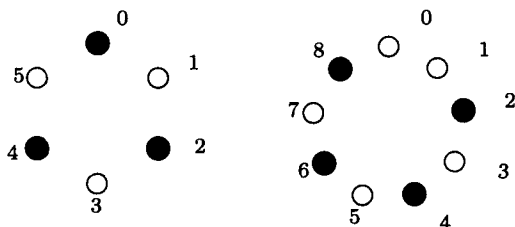


Figure 1.17. The Josephus problem for $n = 5, 8$.

1.38 Example (Sum of the squares of the first n naturals). There is a closed formula for

$$S_2(n) := 1 + 4 + 9 + \cdots + n^2 = \sum_{j=1}^n j^2.$$

In fact the method (iii) in Example 1.36 extends to the present case. Using the identity $(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$, we write

$$\begin{array}{rccccccc} 1^3 & = & & + & & + & & + & 1 \\ 2^3 & = & 1^3 & + & 3 \cdot 1^2 & + & 3 \cdot 1 & + & 1 \\ 3^3 & = & 2^3 & + & 3 \cdot 2^2 & + & 3 \cdot 2 & + & 1 \\ \dots & & & & & & & & \\ n^3 & = & (n-1)^3 & + & 3 \cdot (n-1)^2 & + & 3 \cdot (n-1) & + & 1 \\ (n+1)^3 & = & n^3 & + & 3 \cdot n^2 & + & 3 \cdot n & + & 1. \end{array}$$

Summing, we then get

$$1^3 + 2^3 + \cdots + n^3 + (n+1)^3 = 1^3 + 2^3 + \cdots + n^3 + 3S_2(n) + 3S_1(n) + (n+1),$$

that is

$$3S_2(n) = (n+1)^3 - (n+1) - 3S_1(n),$$

i.e., because of the value of $S_1(n)$ in Example 1.36,

$$\sum_{j=1}^n j^2 = \frac{n(n+1)(2n+1)}{6}, \quad \forall n \geq 1.$$

We can also prove it by induction observing that the sequence $x_n := n(n+1)(2n+1)/6$, $n \geq 1$, satisfies the recursion

$$\begin{cases} x_1 = 1, \\ x_{n+1} = x_n + (n+1)^2 \quad \forall n \geq 1 \end{cases}$$

that defines $S_2(n)$.

1.39 Example (The Josephus problem). Consider the following variant of a story told by Flavius Josephus, a Jewish historian of the first century. We consider n people numbered from 0 to $n-1$, and, starting with the person labelled 1, we eliminate every *second* remaining person until only one survives, see Figure 1.17. We are asked to determine the position $T(n)$ of the survivor.

We easily see that $T(1) = 0$, $T(2) = 0$, $T(3) = 2$, $T(4) = 0$. For large n we may argue recursively. If the number of people is even, after the first round only the even-numbered people survive and the next to be eliminated is labelled 2. We are therefore in the situation of p people numbered $0, 2, \dots, 2p-2$. In formula,

$$T(2p) = 2T(p).$$

If $n = 2p + 1$ is odd, after one round p people labelled $2, 4, \dots, 2p - 2, 2p$ survive and the next to be eliminated is person number 2. Hence

$$T(2p + 1) = 2T(p) + 2.$$

Therefore the sequence of the $T(n)$ satisfies the recurrence

$$\begin{cases} T(1) = 0, \\ T(2n) = 2T(n), & \forall n \geq 0, \\ T(2n + 1) = 2T(n) + 2, & \forall n \geq 0. \end{cases} \quad (1.4)$$

This is the table of $T(n)$ for $n = 0, 1, \dots, 16$.

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$T(n)$	0	0	2	0	2	4	6	0	2	4	6	8	10	12	14	0

This suggests for $T(n)$ the closed form

$$T(n) = 2(n - 2^k), \quad \text{if } 2^k \leq n < 2^{k+1}. \quad (1.5)$$

This is in fact the case as one proves, checking that the sequence $\{x_n\}$ in (1.5) satisfies the recurrence (1.4).

1.4 Summing Up

Real Numbers

The system of real numbers \mathbb{R} is defined axiomatically as a set of objects satisfying a suitable family of rules. First, we can operate on it with addition, multiplication and order in the usual way: this is summarized by saying that \mathbb{R} is an *ordered field*. Secondly, a “continuity property,” that can be expressed in several equivalent forms, holds.

- Let $A \subset \mathbb{R}$ be nonempty. The *supremum* of A , denoted by $\sup A$, is the least upper bound of A , that is, the unique number $L \in \mathbb{R}$ such that
 - (i) L is an upper bound of A , i.e., $x \leq L \forall x \in A$,
 - (ii) $\forall \epsilon > 0$ $L - \epsilon$ is not an upper bound of A , i.e., $\forall \epsilon > 0 \exists x \in A$ such that $x > L - \epsilon$.
- Let $A \subset \mathbb{R}$ be nonempty. The *infimum* of A , denoted by $\inf A$, is the greatest lower bound of A , that is, the unique number $\ell \in \mathbb{R}$ such that
 - (i) ℓ is a lower bound of A , i.e., $x \geq \ell \forall x \in A$,
 - (ii) $\forall \epsilon > 0 \ell + \epsilon$ is not an upper bound of A , i.e., $\forall \epsilon > 0 \exists x \in A$ such that $x < \ell + \epsilon$.
- A *cut* (A, B) of \mathbb{R} is a subdivision of \mathbb{R} in nonempty subsets A and B such that $A \cup B = \mathbb{R}$, $A \cap B = \emptyset$ and

$$\forall a \in A \text{ and } \forall b \in B \text{ we have } a < b.$$

If (A, B) is a cut of X , we say that $x \in X$ corresponds to (A, B) if $a \leq x \leq b \forall a \in A, \forall b \in B$.

The axiom of continuity of the reals can be expressed by one of the following equivalent statements:

- every nonempty subset $A \subset \mathbb{R}$ that is bounded above has supremum, $\sup A \in \mathbb{R}$,
- every nonempty subset $A \subset \mathbb{R}$ that is bounded below has infimum, $\inf A \in \mathbb{R}$,
- to every cut (A, B) of \mathbb{R} corresponds an element of \mathbb{R} .

Natural Numbers

A set $A \subset \mathbb{R}$ is *inductive* if $0 \in A$ and, if $x \in A$, then $x + 1 \in A$. The set of natural numbers is the subset of \mathbb{R} defined by

- \mathbb{N} is the smallest inductive subset of \mathbb{R} .

Relevant facts about naturals are the following:

- INDUCTION PRINCIPLE If $A \subset \mathbb{N}$ is inductive, then $A = \mathbb{N}$.
- INDUCTION PRINCIPLE Suppose that for every natural number $n \in \mathbb{N}$ we are given a statement $p(n)$ and let $k \in \mathbb{N}$.
 - (i) Suppose that the statement $p(k)$ is known to be true.
 - (ii) Suppose that for any $n \geq k$, if the statement $p(n)$ happens to be true, then the statement $p(n + 1)$ must also be true.

Then the statement $p(n)$ must be true for all $n \geq k$.

- ARCHIMEDEAN PROPERTY \mathbb{N} is not bounded above in \mathbb{R} ,
- every nonempty subset $A \subset \mathbb{N}$ has a minimum,
- a subset $A \subset \mathbb{N}$ is bounded if and only if it has a maximum.

Rationals

The integral numbers \mathbb{Z} and the rational numbers \mathbb{Q} are defined respectively by

$$\mathbb{Z} := \{x \in \mathbb{R} \mid |x| \in \mathbb{N}\}, \quad \mathbb{Q} := \left\{x \in \mathbb{R} \mid x = \frac{p}{q}, p, q \in \mathbb{Z}, q \neq 0\right\}.$$

- \mathbb{Q} and the irrationals $\mathbb{R} \setminus \mathbb{Q}$ are dense in \mathbb{R} .

1.5 Exercises

1.40 ¶. Show that $\mathbb{R}_+ := \{x \in \mathbb{R} \mid x > 0\}$ is a multiplicative group. Establish an isomorphism of groups between the additive group \mathbb{R} and the multiplicative group \mathbb{R}_+ .

1.41 ¶. Let X be an ordered field. Show

Proposition. Let $A \subset X$. A has a maximum if and only if A is nonempty, bounded above, has supremum and $\sup A \in A$. In this case $\max A = \sup A$.

1.42 ¶. Show that

- (i) $\inf A \leq \sup A$.
- (ii) If $\emptyset \neq A \subset B \subset \mathbb{R}$, then $\inf B \leq \inf A \leq \sup A \leq \sup B$.
- (iii) Let $A, B \subset \mathbb{R}$ be such that $a \leq b$ for all $a \in A, b \in B$. Then $\inf A \leq \inf B$ and $\sup A \leq \sup B$.

1.43 ¶. Given $A, B \subset \mathbb{R}$ and $\gamma \in \mathbb{R}, \gamma > 0$, define

$$\begin{aligned} A + B &:= \{x \in \mathbb{R} \mid x = a + b, a \in A, b \in B\}, \\ A \cdot B &:= \{x \in \mathbb{R} \mid x = ab, a \in A, b \in B\}, \\ \gamma A &:= \{x \in \mathbb{R} \mid x = \gamma a, a \in A\}, \\ -A &:= \{x \in \mathbb{R} \mid -x \in A\}. \end{aligned}$$

Show that

$$\sup(A + B) = \sup A + \sup B, \quad \inf(A + B) = \inf A + \inf B.$$

If moreover $A, B \subset \mathbb{R}_+$, then

$$\begin{aligned} \sup(A \cdot B) &= \sup A \cdot \sup B, & \inf(A \cdot B) &= \inf A \cdot \inf B, \\ \sup(\gamma A) &= \gamma \sup A, & \inf(\gamma A) &= \gamma \inf A, \\ \inf(-A) &= -\sup A, & \sup -A &= \inf A. \end{aligned}$$

1.44 ¶. Let X be an ordered field. Show that $\min A = -\max(-A)$, $\sup A = -\inf(-A)$ for all $A \subset X$. Deduce that the axioms (C) and (C₁) are equivalent.

1.45 ¶¶. Prove

Proposition. Let $X \subset \mathbb{R}$ be such that

- (i) $0 \in X$,
- (ii) if $x \in X$, then $x + 1 \in X$,
- (iii) if $0 < x \in X$, then $x - 1 \in X$,
- (iv) every nonempty subset A of X has a minimum.

Then $X = \mathbb{N}$.

[Hint: The statements (i) and (ii) say that X is inductive, hence $\mathbb{N} \subset X$. Assume $X \setminus \mathbb{N}$ is nonempty]

1.46 ¶. Theorem 1.28 says that between two distinct real numbers there is a rational one. Show that actually there are infinite many rationals.

1.47 ¶. Show that *irrational numbers* $\mathbb{R} \setminus \mathbb{Q}$ are dense in \mathbb{R} . [Hint: Proceed similarly to the proof of Theorem 1.28.]

1.48 ¶. Show that $2 + \sqrt{3}$ and $\sqrt{2} + \sqrt{3}$ are irrational numbers.

1.49 ¶. Given four rational numbers a, b, c and d with $ad - bc \neq 0$ and an irrational number x such that $cx + d \neq 0$, show that $\frac{ax+b}{cx+d}$ is an irrational number.

1.50 ¶. Let m, n be natural numbers with \sqrt{m} irrational. Show that $\sqrt{m} + \sqrt[k]{n}$ is irrational for all $k \in \mathbb{N}$.

1.51 ¶. Show that, if $a + b\sqrt{2} + c\sqrt[3]{4} = 0$, then $a = b = c = 0$.

1.52 ¶. Show that $\log_{10} 2$ is irrational.

1.53 ¶. Show that the set of rationals $B := \{q \in \mathbb{Q} \mid q^2 \leq 2\}$ has supremum (in \mathbb{R}) given by $\sqrt{2}$.

1.54 ¶ Tarski's paradox. All numbers are equal. We proceed by induction on the number of numbers. The claim is trivial for one number a , $a = a$. Suppose that the claim is true for 3 numbers and let us prove it for 4 numbers, a, b, c, d . We know that $a = b = c$ and $b = c = d$ hence $a = b = c = d$. By induction the claim is proved. Where is the error?

1.55 ¶. Show Proposition 1.32.

1.56 ¶. Show that

$$\begin{aligned} n! &\geq 2^n \quad \forall n \geq 4, \\ 2^n - n &\geq n^2 \quad \forall n \geq 5, \\ n^n &\geq n! \quad \forall n \geq 1. \end{aligned}$$

1.57 ¶ Ovals. Ovals are boundaries of convex figures in the plane. Draw in the plane n ovals. Suppose that each one intersects any other in exactly two points and that no more than three ovals meet at the same point. In how many regions is the plane divided by the ovals?

1.58 ¶. Let I be an interval and let $\phi : I \rightarrow \mathbb{R}$ be convex. Show by induction the *discrete Jensen inequality*, compare Proposition 5.62 of [GM1]:

$$\phi \left(\sum_{i=1}^n \lambda_i x_i \right) \leq \sum_{i=1}^n \lambda_i \phi(x_i) \quad (1.6)$$

for all nonnegative $\lambda_1, \lambda_2, \dots, \lambda_n$ with $\sum_{i=1}^n \lambda_i = 1$ and all $x_1, x_2, \dots, x_n \in I$.

1.59 ¶ Lagrange's identity. Show that

$$\left(\sum_{i=1}^n a_i^2 \right) \left(\sum_{i=1}^n b_i^2 \right) - \left(\sum_{i=1}^n a_i b_i \right)^2 = \sum_{i,j} (a_i b_j - a_j b_i)^2.$$

1.60 ¶. Given reals $\lambda_1, \dots, \lambda_n$ with $0 \leq \lambda_i \leq 1$, show that $\prod_{i=1}^n (1 - \lambda_i) \geq 1 - \prod_{i=1}^n \lambda_i$.

1.61 ¶¶. Show that $n^{n/2} \leq n! \leq ((n+1)/2)^n$. [Hint: Show that $n!^2 = \prod_{i=1}^n k(n+1-k)$ and that for all k , $1 \leq k \leq n$, we have $n \leq k(n+1-k) \leq \frac{1}{4}(n+1)^2$.]

1.62 ¶¶. Let R be a rotation of the plane around the origin of an angle α incommensurable with π . Denote R^n the composition of R with itself, $R^n = R \circ R \circ R \circ \dots \circ R$, n -times. Given a point θ on the unit circle, show that the *orbit* of θ , i.e.,

$$\{z \in \mathbb{R}^2 \mid z = R^n \theta, n \in \mathbb{N}\},$$

is dense in the circle.

1.63 ¶¶. Let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. Show that $\{m\alpha - n \mid m, n \in \mathbb{N}\}$ is dense in \mathbb{R} . Deduce that $\{\sin n \mid n \in \mathbb{N}\}$ is dense in $[-1, 1]$.

1.64 ¶ Galileo. Show that

$$\frac{1}{3} = \frac{1+3}{5+7} = \frac{1+3+5}{7+9+11} = \dots$$

1.65 ¶. Find closed formulas for $\sum_{j=0}^n j^2 q^j$, $\sum_{j=0}^n j^3 q^j$.

1.66 ¶. Compute

$$\begin{aligned} \sum_{j=1}^n (j-1)^2, & \qquad \sum_{j=1}^n (j^2 - j + 1), \forall n \geq 1, \\ \sum_{j=1}^n j(j-1), & \qquad \sum_{j=1}^n j(j+1)(j+2), \\ \sum_{j=1}^n j(j+1)(j+2)(j+3). & \end{aligned}$$

1.67 ¶ Nicomachus's theorem. Show that

$$\sum_{j=1}^n j^3 = \left(\sum_{j=1}^n j \right)^2, \quad \forall n \geq 1.$$

1.68 ¶ Catalan's identity. Show that

$$\sum_{j=1}^n \frac{1}{n+j} = \sum_{j=1}^{2n} (-1)^{n-j} \frac{1}{j}, \quad \forall n \geq 1.$$

1.69 ¶. n straight lines are said to be in a generic position if they intersect each other at one and only one point. Determine how many regions are delimited by n straight lines in generic position in the plane.

1.70 ¶. Show that $\sum_{j=0}^n \binom{n}{j} = 2^n$.

Thales of Miletus (624BC–546BC)		Pythagoras of Samos (580BC–520BC)		
Hippocrates of Chios (470BC–410BC)	Hippias of Elis (460BC–400BC)	Plato (428BC–347BC)	Aristotle (384BC–322BC)	Eudemus of Rhodes (350BC–290BC)
Eudoxus of Cnidus (408BC–355BC)				
Euclid of Alexandria (325BC–265BC)				
Aristarchus of Samos (310BC–230BC)	Eratosthenes of Cyrene (276BC–197BC)		Nicomedes (280BC–210BC)	
Archimedes of Syracuse (287BC–212BC)				
Apollonius of Perga (262BC–190BC)				
Diocles (240BC–180BC)	Hipparchus of Rhodes (190BC–120BC)	Zenodorus (200BC–140BC)	Menelaus of Alexandria (70AD–130)	
Ptolemy (85–165)				
Heron of Alexandria (IAD)		Nicomachus of Gerasa (60AD–120)		
Diophantus of Alexandria (200–284)				
Pappus of Alexandria (290–350)	Theon of Alexandria (335–395)	Diadochus Proclus (411–485)	Anicus Boethius (475–524)	Eutocius of Ascalon (480–540)

Figure 1.18. A table of Greek Mathematicians.

2. Sequences of Real Numbers

As we have seen, we can represent any rational number, for instance $\sqrt{2}$, by its successive approximations with rational numbers, q_1, q_2, \dots . According to Greek mathematicians the process which generates the approximations q_1, q_2, \dots never ends; for us, instead, such a process is the realization of $\sqrt{2}$ as the *limit of the sequence* $\{q_n\}$. In this chapter we shall discuss the notions of sequence and of limit of a sequence.

In Section 2.1 we discuss basic properties. They may be inferred by analogous properties for limits of functions proved in [GM1]. However, we supply direct proofs for two reasons: first to be self-contained and, secondly, because one may want to discuss limits of sequences before limits of functions. In Section 2.2 we discuss the important notion of *Cauchy sequence*, we prove the *Bolzano–Weierstrass theorem* and give various equivalent formulations of continuity of the reals. In Section 2.3 we give alternative simple proofs of the intermediate value and Weierstrass theorems. Finally, in Section 2.4 we discuss a few examples, and in Section 2.5 we give an alternative definition, in terms of sequences, i.e., just continuity, of the exponential and logarithmic functions.

2.1 Sequences

2.1 Definition. A sequence with values in a set X , or simply a sequence in X , is a function $x : \mathbb{N} \rightarrow X$.

A sequence is denoted by $\{x_n\}$, $n \geq 0$, or by $\{x_n\}_{n \in \mathbb{N}}$; x_n , that is $x(n)$, is referred to as to the n -th term of the sequence $\{x_n\}$. Accordingly, any enumeration of points of X by means of an index, which varies in an infinite subset of the integers, is called a sequence, too: for instance we say “the sequence $1/n$ with n odd” for $\{x_n\}_{n \in \mathbb{N}}$ with $x_n = 1/(2n+1)$.

There are many ways to produce sequences. For example we can give a formula to compute x_n for all n , as

$$x_n = \frac{1}{n}, \quad n \geq 1, \quad x_n = \frac{n^2 + \sin(1/n)}{n!}, \quad n \geq 1,$$

or, and this is in many respects more interesting as we shall see later in Chapter 8, we can give a rule which gives each term of the sequence in terms of the preceding terms as in

$$\begin{cases} x_0 = 1, \\ x_{n+1} = f(x_n), \quad \forall n \geq 0, \end{cases} \quad (2.1)$$

where $f : \mathbb{R} \rightarrow \mathbb{R}$ is a given function. In this case we can compute

$$\begin{aligned} x_0 &= 1, \\ x_1 &= f(x_0) = f(1), \\ x_2 &= f(x_1) = f(f(1)), \\ &\dots \end{aligned}$$

and it appears clearly that (2.1) defines uniquely the sequence $\{x_n\}$. Actually, this is a consequence of the induction principle: the set

$$A := \left\{ n \in \mathbb{N} \mid x_n \text{ is defined by (2.1)} \right\}$$

is inductive, hence $A = \mathbb{N}$, i.e., $\{x_n\}$ is uniquely defined for all $n \in \mathbb{N}$. It is usual to refer to (2.1) as to the *recursive definition* of $\{x_n\}$ or to the *recursive sequence* $\{x_n\}_n$.

2.2 Example (Integer powers). If $q \in \mathbb{R}$ and $n \in \mathbb{N}$, then q^n is defined as the product of q by itself n times,

$$q^n := \underbrace{q q q \cdots q}_n, \quad n \text{ times.}$$

However this is a costly definition: we need to recompute with an increasing number of multiplications every time we increase the exponent. A simpler way to define *all* expressions q^n , $n \in \mathbb{N}$, is by the recursive definition

$$\begin{cases} q^0 = 1, \\ q^{n+1} = q q^n \quad \forall n \geq 0. \end{cases} \quad (2.2)$$

2.3 Example (Products). Let $\{a_n\}$, $n \geq 0$, be a sequence of real numbers. The product of the first n -terms of the sequence is

$$\prod_{i=1}^n a_i := a_1 a_2 \cdots a_n.$$

The sequence $x_n := \prod_{j=0}^n a_j$, $n \geq 0$, is clearly defined recursively by

$$\begin{cases} x_1 := a_1, \\ x_{n+1} := x_n a_{n+1} \quad \forall n \geq 1. \end{cases} \quad (2.3)$$

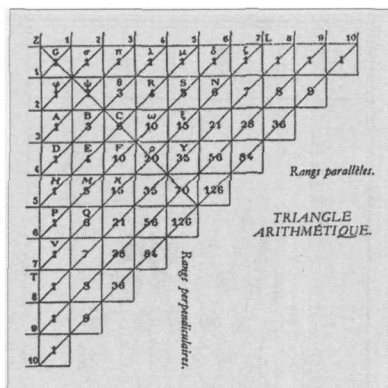


Figure 2.1. Pascal's triangle.

2.4 Example (Factorial). For all $n \in \mathbb{N}$, the *factorial* $n!$ of n is defined as 1 if $n = 0$ and as the product of the first n natural numbers

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

if $n \geq 1$. All factorials are, in fact, defined by

$$\begin{cases} x_0 = 1, \\ x_{n+1} = (n+1)x_n. \end{cases}$$

2.5 Example (Sums). The sum of the first $n+1$ terms of a sequence $\{a_n\}_{n \geq 0}$, of real numbers $a_0 + a_1 + \cdots + a_n$ is denoted by

$$\sum_{j=0}^n a_j.$$

The sequence $x_n := \sum_{j=0}^n a_j$ is defined by

$$\begin{cases} s_0 = a_0, \\ s_{n+1} = s_n + a_{n+1} \quad \forall n \geq 0. \end{cases}$$

In $\sum_{j=0}^n a_j$, the integral variable j , which varies from 0 to n , just enumerates the elements to be summed: it is a *bound variable*: we clearly have

$$\sum_{j=0}^n a_j = \sum_{k=0}^n a_k = \sum_{j+2=0}^n a_{j+2} = \sum_{j=2}^{n+2} a_{j-2}.$$

2.6 Example (Binomial coefficients and Newton's binomial). The binomial coefficients (see, for example, Chapter 4 of [GM1]) are defined by

$$\binom{n}{j} := \frac{n!}{j!(n-j)!} = \frac{n(n-1)(n-2) \cdots (n-j+1)}{j!}, \quad \forall j, 0 \leq j \leq n.$$

It is clearly seen that for $j, n \in \mathbb{N}$ and $0 \leq j \leq n$ we have

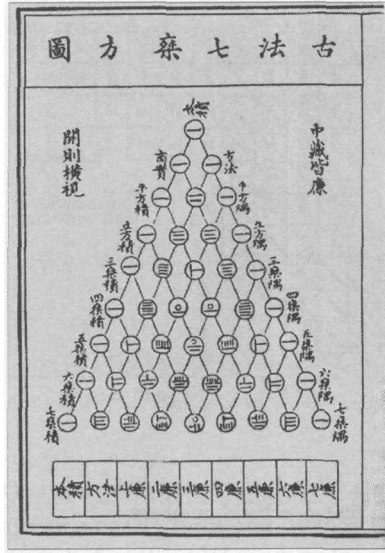


Figure 2.2. Pascal's triangle.

$$\begin{aligned}
 \binom{n}{0} &= \binom{n}{n} = 1, & \binom{n}{1} &= \binom{n}{n-1} = n, \\
 \binom{n}{j} &= \binom{n}{n-j}, & \binom{n}{j} &= \frac{n}{j} \binom{n-1}{j-1}, \\
 \binom{n}{k} &= \binom{n-1}{k-1} + \binom{n-1}{k}, \quad 1 \leq k \leq n,
 \end{aligned}$$

where the last formula is known as *Pascal's formula*.

As an application of the induction principle let us give a proof of the *binomial theorem*

$$(a+b)^n = \sum_{k=0}^n \binom{n}{k} a^{n-k} b^k, \quad \forall a, b \in \mathbb{R} \quad (2.4)$$

(see, for example, 5.53 of [GM1]), which makes no use of calculus.

The claim (2.4) is trivial if either a or b is zero. If, say, a is nonzero, by multiplying and dividing by a^n , we see that (2.4) is equivalent to

$$(1+h)^n = \sum_{k=0}^n \binom{n}{k} h^k. \quad (2.5)$$

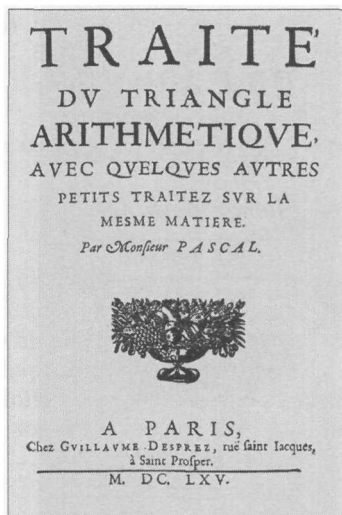
Therefore it suffices to show that the sequence $x_n := \sum_{j=0}^n \binom{n}{j} h^j$ satisfies the same recursive definition of $(1+h)^n$, i.e.,

$$\begin{cases} x_0 = 1, \\ x_{n+1} = (1+h)x_n \quad \forall n \geq 0. \end{cases}$$

Since in fact $x_0 = \sum_{j=0}^0 \binom{0}{j} h^j = 1$ and



Figure 2.3. Blaise Pascal (1623–1662) and the frontispiece of his *Traité du triangle arithmétique*.



$$\begin{aligned}
 x_n(1+h) &= \sum_{j=0}^n \binom{n}{j} h^j + \sum_{j=0}^n \binom{n}{j} h^{j+1} = \sum_{j=0}^n \binom{n}{j} h^j + \sum_{j=1}^{n+1} \binom{n}{j-1} h^j \\
 &= 1 + \sum_{j=1}^n \left\{ \binom{n}{j} + \binom{n}{j-1} \right\} h^j + h^{n+1} = 1 + \sum_{j=1}^n \binom{n+1}{j} h^j + h^{n+1} = x_{n+1},
 \end{aligned}$$

on account of Pascal's formula, the claim is proved.

a. Limit of a sequence

The notion of limit of a sequence plays a fundamental role in analysis.

2.7 Definition (of limit). Let $\{x_n\}$ be a sequence of real numbers and let $L \in \mathbb{R}$. We say that $\{x_n\}$ tends to , or converges to L , or that L is the limit of $\{x_n\}$, and we write

$$x_n \rightarrow L \quad \text{or} \quad \lim_{n \rightarrow \infty} x_n = L$$

if

$$\forall \epsilon > 0 \exists n_\epsilon \in \mathbb{N} \text{ such that } |x_n - L| < \epsilon \forall n \geq n_\epsilon. \quad (2.6)$$

2.8 Proposition. $1/n \rightarrow 0$ and $(-1)^n/n \rightarrow 0$.

2.9 ¶. Show that that $x_n \rightarrow L$ if and only if $|x_n - L| \rightarrow 0$.

2.10 ¶. Show that the two claims in Proposition 2.8 are both equivalent to the Archimedean property.

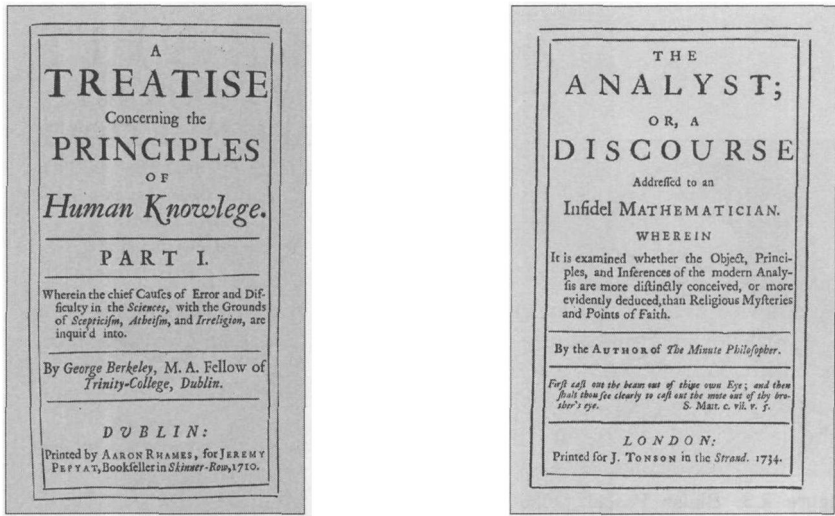


Figure 2.4. Frontispieces of *A Treatise concerning the principles of human knowledge* and of *The Analyst* by George Berkeley (1685–1753).

2.11 Definition. We say that $\{x_n\}$ diverges or tends to $+\infty$, or that $+\infty$ is the limit of $\{x_n\}$, and we write

$$x_n \rightarrow +\infty \quad \text{or} \quad \lim_{n \rightarrow \infty} x_n = +\infty$$

if

$$\forall M > 0 \exists \nu \in \mathbb{N} \text{ such that } x_n > M \quad \forall n \geq \nu.$$

We say that $\{x_n\}$ diverges or tends to $-\infty$, or that $-\infty$ is the limit of $\{x_n\}$, and we write

$$x_n \rightarrow -\infty \quad \text{or} \quad \lim_{n \rightarrow \infty} x_n = -\infty$$

if

$$\forall M > 0 \exists \nu \in \mathbb{N} \text{ such that } x_n < -M \quad \forall n \geq \nu.$$

Finally we say that $\{x_n\}$ has a limit if $\{x_n\}$ converges or diverges.

b. Properties of limits and calculus

We may interpret the limit of a sequence as the limit of a function. In fact, given $\{a_n\}$, fix an interval $[x_0, q[$, $q \in \mathbb{R}$ (say $[0, +\infty[$), and a strictly increasing sequence $\{x_n\}$ in $[x_0, q[$ with $x_n \rightarrow q$ (say $x_n = n$ if $q = +\infty$), and define the step function $\varphi_a : [x_0, q[\rightarrow \mathbb{R}$ by

$$\varphi_a(x) = a_n \quad \text{if} \quad x_n \leq x < x_{n+1}.$$

2.12 Proposition. $a_n \rightarrow L \in \overline{\mathbb{R}}$ if and only if $\varphi_a(x) \rightarrow L$ as $x \rightarrow q^-$.

2.13 ¶. Prove Proposition 2.12.

2.14 ¶. Let $\{a_n\}, \{b_n\}$ be two sequences, $\{x_n\}$ as above, and let φ_a, φ_b be the corresponding functions. Show that

- $\varphi_{a+b}(x) = \varphi_a(x) + \varphi_b(x)$,
- $\varphi_{ab}(x) = \varphi_a(x)\varphi_b(x)$.

Proposition 2.12 and Exercise 2.14 allow us to specialize properties and results we have already proved for limits of functions to limits of sequences. We however add the simple direct proofs, which are similar to the ones of the limits of functions (see, for example, Chapter 2 of [GM1]). In fact one might want to develop first the theory of limits of sequences and then the theory of limits of functions. As we shall see in Theorem 2.46 the two approaches are completely equivalent.

2.15 Proposition. We have:

- (i) (UNIQUENESS) A sequence cannot have more than one limit.
- (ii) (BOUNDEDNESS) If $\{x_n\}$ converges, then $\{x_n\}$ is bounded.
- (iii) (CONSTANCY OF SIGN) Suppose that $\{x_n\}$ has limit $L \in \mathbb{R}$.
 - a) If $L > 0$ (respectively $L < 0$), then there exists \bar{n} such that $x_n > 0$ (respectively $x_n < 0$) for all $n \geq \bar{n}$.
 - b) If there exists \bar{n} such that $x_n \geq 0$ (respectively $x_n \leq 0$) for all $n \geq \bar{n}$, then $L \geq 0$ (respectively $L \leq 0$).

Proof. (i) Suppose $x_n \rightarrow L_1$, $x_n \rightarrow L_2$, and $L_1 \neq L_2$. If $L_1, L_2 \in \mathbb{R}$, then for $\epsilon := |L_1 - L_2|/2$ we find x_ν such that $|x_\nu - L_1| < \epsilon$ and $|x_\nu - L_2| < \epsilon$. Therefore

$$2\epsilon = |L_1 - L_2| \leq |L_1 - x_\nu| + |x_\nu - L_2| < 2\epsilon,$$

a contradiction. The cases in which L_1 and/or L_2 are infinity are similar.

(ii) Let $\{x_n\}$ converge to L . By definition, choosing $\epsilon = 1$, we find \bar{n} such that $|x_n - L| < 1$ for all $n \geq \bar{n}$. In particular $|x_n| \leq |x_n - L| + |L| < 1 + |L|$ for $n \geq \bar{n}$. Hence

$$M := |x_1| + |x_2| + \cdots + |x_{\bar{n}-1}| + |L| + 1$$

is an upper bound for $\{|x_n|\}$.

(iii) Suppose $L > 0$. From the definition of limit with $\epsilon = L/2$, we find $\bar{n} \in \mathbb{N}$ such that $|x_n - L| < L/2$ for all $n \geq \bar{n}$, that is, $0 < L/2 = L - L/2 < x_n < 3L/2$, which proves (a). By contradiction one then sees that (b) is equivalent to (a). \square

2.16 ¶ Sequences need not have limits. Show that $x_n := (-1)^n$ has no limit as $n \rightarrow \infty$.

2.17 ¶. Show that, if $x_n \rightarrow L$ and $x_n > 0 \forall n$, then L need not be positive.

2.18 Proposition (Squeezing and comparison test). Let $\{a_n\}, \{b_n\}$ and $\{c_n\}$ be three sequences. Suppose that there exists \bar{n} such that $a_n \leq b_n \leq c_n \forall n \geq \bar{n}$. If $a_n \rightarrow L$ and $c_n \rightarrow L$, then $b_n \rightarrow L$.

Proof. Suppose $L \in \mathbb{R}$; we leave to the reader the discussion of the cases $L = \pm\infty$. Given $\epsilon > 0$ we find n_ϵ such that $L - \epsilon < a_n < L + \epsilon$ and $L - \epsilon < c_n < L + \epsilon$. Since $a_n \leq b_n \leq c_n$ for all $n \geq \bar{n}$, we conclude that $L - \epsilon < b_n < L + \epsilon$ for all $n \geq \max(\bar{n}, n_\epsilon)$, that is $b_n \rightarrow L$. \square

In the applications, Proposition 2.18 is often used in the following version.

2.19 Corollary. *Let $\{x_n\}$, $\{y_n\}$ be two sequences and let $L \in \mathbb{R}$. If*

$$\exists \bar{n} \text{ such that } |x_n - L| \leq y_n, \quad \forall n \geq \bar{n}, \quad \text{and} \quad y_n \rightarrow 0,$$

then $x_n \rightarrow L$. If

$$\exists \bar{n} \text{ such that } x_n \leq y_n \quad \forall n \geq \bar{n}, \quad \text{and} \quad x_n \rightarrow +\infty,$$

then $y_n \rightarrow +\infty$.

2.20 Proposition. *Suppose $\{x_n\}$ and $\{y_n\}$ have limits respectively ℓ and m in $\overline{\mathbb{R}}$. Then*

- (i) *If $\ell + m$ is well defined in $\overline{\mathbb{R}}$, then $x_n + y_n \rightarrow \ell + m$.*
- (ii) *If ℓm is well defined in $\overline{\mathbb{R}}$, then $x_n y_n \rightarrow \ell m$.*
- (iii) *If ℓ/m is well defined in $\overline{\mathbb{R}}$, then $x_n/y_n \rightarrow \ell/m$.*
- (iv) *If $y_n \rightarrow 0$ and $y_n > 0$ for all n , then $1/y_n \rightarrow +\infty$.*

Proof. We prove (i) in the case $\ell, m \in \mathbb{R}$. The reader is asked to discuss the other cases. Given $\epsilon > 0$ we find n_x and n_y such that

$$|x_n - \ell| < \epsilon \text{ for all } n \geq n_x \quad |y_n - m| < \epsilon \text{ for all } n \geq n_y,$$

hence for $n \geq \bar{n} := \max(n_x, n_y)$, the two inequalities $|x_n - \ell| < \epsilon$ and $|y_n - m| < \epsilon$ hold. By the triangle inequality we then infer

$$|x_n - \ell + y_n - m| \leq |x_n - \ell| + |y_n - m| < \epsilon + \epsilon = 2\epsilon \quad \text{for all } n \geq \bar{n},$$

which yields the conclusion, since ϵ is arbitrary.

(ii) If $\ell, m \in \mathbb{R}$ we write

$$|x_n y_n - \ell m| = |x_n(y_n - m) + m(x_n - \ell)| \leq |x_n| |y_n - m| + |m| |x_n - \ell| \leq (K + |m|)\epsilon$$

where K is an upper bound for $\{|x_n|\}$ (see, for example, Proposition 2.15). This yields the conclusion. The cases $\ell = \pm\infty$ and $m = \pm\infty$ or $m \neq 0$, are simpler.

(iii) If $\ell, m \in \mathbb{R}$, $m \neq 0$, it suffices to notice that

$$\left| \frac{x_n}{y_n} - \frac{\ell}{m} \right| = \left| \frac{mx_n - \ell m + \ell m - \ell y_n}{my_n} \right| \leq \frac{1}{|y_n|} \left(|x_n - \ell| + \frac{|\ell|}{|m|} |y_n - m| \right)$$

to conclude the proof. (iv) We ask the reader to prove it. \square

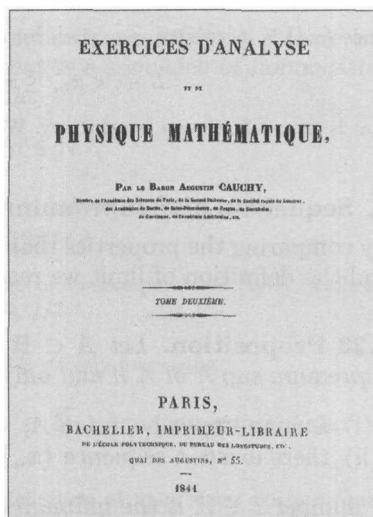


Figure 2.5. The frontispiece of *Exercices d'Analyse et de Physique Mathématique* by Augustin-Louis Cauchy (1789–1857).

c. Limits of monotone sequences

2.21 Definition. A sequence $\{x_n\}$ of real numbers is said to be

- bounded above if $\exists c \in \mathbb{R}$ such that $x_n \leq c \ \forall n \in \mathbb{N}$,
- bounded below if $\exists c \in \mathbb{R}$ such that $x_n \geq c \ \forall n \in \mathbb{N}$,
- bounded if $\exists c \in \mathbb{R}$ such that $|x_n| \leq c \ \forall n \in \mathbb{N}$,
- increasing if $\forall n$ we have $x_n \leq x_{n+1}$,
- decreasing if $\forall n$ we have $x_n \geq x_{n+1}$,
- strictly increasing if $\forall n$ we have $x_n < x_{n+1}$,
- strictly decreasing if $\forall n$ we have $x_n > x_{n+1}$,
- monotone if it is increasing or decreasing,
- strictly monotone if it is strictly increasing or strictly decreasing.

Recall that we write

$$\sup A = +\infty \quad (\text{respectively } \inf A = -\infty)$$

if A is not bounded above (respectively below). An important consequence of the continuity of the reals, on account of Proposition 2.12 above and of Proposition 2.30 of [GM1] or directly, is

2.22 Proposition (Limits of monotone sequences). *All monotonic sequences have limits. More precisely, if $\{x_n\}$ is increasing, then $x_n \rightarrow \sup_n \{x_n\}$, while if $\{x_n\}$ is decreasing, $x_n \rightarrow \inf_n \{x_n\}$.*

Proof. Suppose $\{x_n\}$ is increasing, and let $L := \sup_n \{x_n\}$, that we assume to be a real minimizer. Given $\epsilon > 0$, the properties of the supremum read

- (i) $x_n \leq L, \ \forall n$,
- (ii) $\exists n_\epsilon$ such that $L - \epsilon < x_{n_\epsilon}$.

Since $\{x_n\}$ is increasing, $x_{n_\epsilon} \leq x_n$ for all $n \geq n_\epsilon$, hence

$$L - \epsilon < x_{n_\epsilon} \leq x_n \leq L \quad \text{for all } n \geq n_\epsilon,$$

that is, $x_n \rightarrow L$, since ϵ is arbitrary. We ask the reader to prove the other cases. \square

d. Sequences and supremum

By comparing the properties that characterize the supremum (or infimum), and the definition of limit we readily see

2.23 Proposition. *Let $A \subset \mathbb{R}$ be nonempty. A number $L \in \mathbb{R}$ is the supremum $\sup A$ of A if and only if*

- (i) *L is an upper bound of A ,*
- (ii) *there exists a sequence $\{x_n\} \subset A$ that converges to L .*

A number $L \in \mathbb{R}$ is the infimum of A if and only if

- (i) *L is a lower bound of A ,*
- (ii) *there exists a sequence $\{x_n\} \subset A$ that converges to L .*

Notice that $\sup A = +\infty$ if and only if there exists a sequence $\{x_n\}$ with values in A that diverges to $+\infty$, in fact $\sup A = +\infty$ is equivalent to the unboundedness of A , i.e., to

$$\forall n > 0 \exists x_n \in A \text{ such that } x_n > n.$$

Similarly $\inf A = -\infty$ if and only if there is a sequence of points in A that diverges to $-\infty$. In conclusion, regardless of boundedness of A , i.e., if A is nonempty, we can always claim the existence of a *maximizing sequence*, i.e., a sequence $\{x_n\} \subset A$ that tends to $\sup A$, and of a *minimizing sequence*, i.e., a sequence $\{x_n\}$ that tends to $\inf A$.

2.24 ¶. More precisely, prove the following two propositions:

Proposition. *Let A be a nonempty subset of \mathbb{R} . Then there exists an increasing sequence $\{x_n\} \subset A$ that converges to $\sup A$. Moreover we can choose $\{x_n\}$ to be strictly increasing if A has no maximum or to be constant if A has maximum.*

Proposition. *Every real number is the limit of a monotone sequence of rational numbers.*

e. Subsequences

Of particular relevance is the notion of *subsequence* of a sequence. If $\{x_n\}$ is a sequence, a subsequence of $\{x_n\}$ is a new sequence obtained by choosing its values among the values of $\{x_n\}$, however not randomly, but keeping a strict order on the indices.

2.25 Definition. We say that $\{y_n\}$ is a subsequence of $\{x_n\}$ if there is a function $k : \mathbb{N} \rightarrow \mathbb{N}$ strictly increasing, that is a sequence of nonnegative integers with $k_1 < k_2 < k_3 < \dots$, such that

$$y_n = x_{k_n} \quad \forall n \in \mathbb{N}.$$

2.26 Example. The first terms of the sequence $\{1/n\}$ are given by

$$1, 1/2, 1/3, 1/4, 1/5, 1/6, 1/7, \dots$$

The terms $1, 1/3, 1/4, 1/7$ may be the first terms of a subsequence of $\{1/n\}$, while no subsequence of $\{1/n\}$ can start with $1, 1/3, 1/4, 1/2, \dots$

2.27 ¶. Notice that $k_n \geq n \quad \forall n$, since $k : \mathbb{N} \rightarrow \mathbb{N}$ is strictly increasing.

2.28 Proposition. If $\{x_n\}$ has limit $L \in \overline{\mathbb{R}}$, then any subsequence of $\{x_n\}$ has the same limit L .

Proof. Suppose $L \in \mathbb{R}$, and let $\epsilon > 0$. By the definition of limit there exists n_ϵ such that $|x_n - L| < \epsilon$ for all $n \geq n_\epsilon$. Since $k_n \geq n$, we also have $|x_{k_n} - L| < \epsilon$ for all $n \geq n_\epsilon$, i.e., $x_{k_n} \rightarrow L$, ϵ being arbitrary. The proofs in the cases $L = \pm\infty$ are similar. \square

In particular if $\{x_n\}$ has two subsequences with different limits, then $\{x_n\}$ has no limit.

The following two examples, though artificially simple, may serve to illustrate the usefulness of Proposition 2.28.

2.29 Example. The sequence $1/n^2 \rightarrow 0$ converges to zero, since it is a subsequence of $\{1/n\}$ (the selection map being $k_n = n^2$). Similarly $1/2^n \rightarrow 0$.

2.30 Example. $\frac{1}{\sqrt{n}} \rightarrow 0$. In fact, since $\{1/\sqrt{n}\}$ is decreasing it has a limit and $1/\sqrt{n} \rightarrow \ell$, $\ell \in \mathbb{R}$. On the other hand $\{1/n\}$ is a subsequence of $\{1/\sqrt{n}\}$ (the selection map being $k_n = n^2$), hence $1/n \rightarrow \ell$, consequently $\ell = 0$, since the limit is unique.

2.2 Equivalent Formulations of the Continuity Axiom

a. The principle of nested intervals or Cantor's principle

2.31 Theorem (Cantor's intersection theorem). Let $C_n = [a_n, b_n]$ be a sequence of closed intervals of \mathbb{R} such that

$$[a_{n+1}, b_{n+1}] \subset [a_n, b_n] \quad \forall n \in \mathbb{N}.$$

Then there exists at least a point x common to all intervals, $x \in \bigcap_{n=1}^{\infty} C_n$.

Proof. Clearly we have

(i) $a_1 \leq a_2 \leq a_3 \leq \dots$, i.e., the sequence $\{a_n\}$ is increasing.

- (ii) $b_1 \geq b_2 \geq b_3 \geq \cdots$, i.e., the sequence $\{b_n\}$ is decreasing.
 (iii) $a_n \leq b_m$ for all n and m .

Consequently $\{a_n\}$ and $\{b_n\}$ are bounded monotone sequences, by Proposition 2.22 they have finite limits, $a_n \uparrow \ell$, $b_n \downarrow L$, and we have

$$a_n \leq \ell \leq L \leq b_m \quad \forall n, m \in \mathbb{N}.$$

In particular

$$[\ell, L] \subset \bigcap_{n=1}^{\infty} C_n.$$

□

2.32 ¶. Show that in the proof of Theorem 2.31 we have $[\ell, L] = \bigcap_{n=1}^{\infty} C_n$.

b. Cauchy criterion

Except for monotone sequences we cannot state that a sequence converges without involving its limit in advance.

2.33 Definition. We say that $\{x_n\}$ is a Cauchy or fundamental sequence if

$$\forall \epsilon \exists \bar{n} \text{ such that } |x_h - x_k| < \epsilon \quad \forall h, k \geq \bar{n}.$$

To a given sequence $\{x_n\}$, we associate a new sequence $\{d_n\}$ defined by

$$d_n := \sup_{h, k \geq n} |x_h - x_k|. \quad (2.7)$$

Notice that d_k can be understood as the length of the interval spanned by all the elements of the sequence $\{x_n\}$ but the first k . Definition 2.33 yields

2.34 Proposition. A sequence $\{x_n\}$ is a Cauchy sequence if and only if the corresponding sequence $\{d_n\}$ in (2.7) tends to zero.

If $x_n \rightarrow \ell$, then clearly for any $\epsilon > 0$ we have $|x_n - x_m| \leq |x_n - \ell| + |x_m - \ell| < \epsilon$ provided n, m are large enough, i.e., $n, m \geq \bar{n}$ in such a way that $|x_n - \ell| < \epsilon/2 \quad \forall n \geq \bar{n}$. In other words: *every convergent sequence is a Cauchy sequence.*

It is an important fact that the opposite holds true.

2.35 Theorem (Cauchy's criterion). A real sequence is convergent if and only if it is a Cauchy sequence.

Proof. It remains to prove that Cauchy's sequences are convergent. We first show that *Cauchy's sequences are bounded*. In fact, choosing $\epsilon = 1$, we find \bar{n} such that

$$|x_n - x_{\bar{n}}| < 1 \quad \forall n \geq \bar{n}.$$

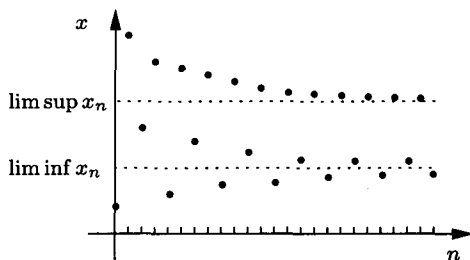


Figure 2.6. \liminf and \limsup .

An upper bound for $\{|x_n|\}$ is then given by $|x_1| + |x_2| + \cdots + |x_n| + 1$.

Define now for $n = 1, 2, 3, \dots$

$$\ell_n := \inf_{k \geq n} \{x_k\}, \quad L_n := \sup_{k \geq n} \{x_k\},$$

and observe that $\{\ell_n\}$, $\{L_n\}$ are sequences of real numbers, $\{x_n\}$ being bounded. Moreover $\{\ell_n\}$ is increasing, $\{L_n\}$ is decreasing and $\ell_n \leq L_n$ $\forall n \in \mathbb{N}$. By Proposition 2.22 we have $\ell_n \uparrow \ell$, $L_n \downarrow L$, and

$$\ell_n \leq \ell \leq L \leq L_n \quad \forall n \in \mathbb{N}.$$

Since $\{x_n\}$ is a Cauchy sequence and

$$L_n - \ell_n = \sup_{h, k \geq n} |x_h - x_k|,$$

we get $L_n - \ell_n \rightarrow 0$ and therefore $\ell = L =: \bar{x} \in \mathbb{R}$. Let us finally prove that $x_n \rightarrow \bar{x}$. Fix $\epsilon > 0$ and let n_ϵ be such that $\bar{x} - \ell_{n_\epsilon}$ and $L_{n_\epsilon} - \bar{x}$ be not greater than ϵ . Since for all $n \geq n_\epsilon$ we have $\ell_{n_\epsilon} \leq x_n \leq L_{n_\epsilon}$, we conclude that

$$\bar{x} - \epsilon \leq \ell_{n_\epsilon} \leq x_n \leq L_{n_\epsilon} \leq \bar{x} + \epsilon,$$

i.e., $x_n \rightarrow \bar{x}$ as $n \rightarrow \infty$, ϵ being arbitrary. \square

2.36 Remark. As we have seen, every real number is the limit of a strictly increasing (respectively decreasing) sequence of rational numbers. Consider the space of all Cauchy sequences of rational numbers in which we identify those Cauchy sequences $\{p_n\}$ and $\{q_n\}$ if $p_n - q_n \rightarrow 0$, i.e., if “they have the same limit.” Cauchy’s criterion then allows us, essentially, to identify this space with \mathbb{R} .

c. Upper and lower limits

Consider any sequence $\{x_n\} \subset \mathbb{R}$. The sequences $\{\ell_n\}$ $\{L_n\}$ previously defined by

$$\ell_n := \inf_{k \geq n} \{x_k\}, \quad L_n := \sup_{k \geq n} \{x_k\}$$

are respectively an increasing sequence and a decreasing sequence of extended real numbers. Consequently they have limits in $\overline{\mathbb{R}}$,

$$\ell_n \rightarrow \sup_k \ell_k, \quad L_n \rightarrow \inf_k L_k.$$

We set

$$\begin{aligned} \liminf_{n \rightarrow \infty} x_n &:= \lim_{n \rightarrow \infty} \ell_n = \lim_{n \rightarrow \infty} \inf_{k \geq n} \{x_k\}, \\ \limsup_{n \rightarrow \infty} x_n &:= \lim_{n \rightarrow \infty} L_n = \lim_{n \rightarrow \infty} \sup_{k \geq n} \{x_k\}, \end{aligned}$$

and refer to them respectively as to the *lower* and *upper limit* or the *limit inferior* and the *limit superior* of $\{x_n\}$. These new notions will be very useful in the sequel, here we confine ourselves to a few comments.

2.37 Proposition. *Every sequence in \mathbb{R} has an upper and lower limit in $\overline{\mathbb{R}}$.*

From the definition

$$\liminf_{n \rightarrow \infty} x_n \leq \limsup_{n \rightarrow \infty} x_n, \quad \text{and} \quad \liminf_{n \rightarrow \infty} x_n = -\limsup_{n \rightarrow \infty} (-x_n).$$

Going to the proof of Cauchy's criterion, we also see

2.38 Proposition. *Let $\{x_n\}$ be a sequence in $\overline{\mathbb{R}}$. Then $x_n \rightarrow \ell \in \overline{\mathbb{R}}$ if and only if*

$$\liminf_n x_n = \limsup_n x_n = \ell.$$

The following proposition characterizes the upper limit of a bounded sequence.

2.39 Proposition. *Let $\{x_n\}$ be a sequence in \mathbb{R} . The number $L \in \mathbb{R}$ is the upper limit of $\{x_n\}$ if and only if*

- (i) $\forall \epsilon > 0 \exists \bar{n}$ such that $x_n < L + \epsilon$ for all $n \geq \bar{n}$,
- (ii) *there exists a subsequence $\{x_{k_n}\}$ of $\{x_n\}$ that converges to L .*

Proof. Let $L = \limsup_n x_n \in \mathbb{R}$ and prove that (i) (ii) hold. By definition $L = \lim_{n \rightarrow \infty} \sup_{k \geq n} \{x_k\}$ hence for $\epsilon > 0$ there is n_ϵ such that

$$L - \epsilon < \sup_{k \geq n} \{x_k\} < L + \epsilon \quad \text{for } n \geq n_\epsilon.$$

Because of the properties of the supremum, the last inequalities hold if and only if

$$x_n \leq L + \epsilon \text{ for } n \geq n_\epsilon, \quad (2.8)$$

$$\text{for all } n \geq n_\epsilon \text{ there is } k \geq n \text{ such that } x_k > L - \epsilon. \quad (2.9)$$

Clearly (2.8) is (i). Let us show a subsequence $\{x_{k_n}\}$ of $\{x_n\}$ converging to L . For $\epsilon = 1$ we choose $n = n_\epsilon$, and, on account of (2.9), we find $k_1 > n_\epsilon$ such that $L - 1 < x_{k_1} < L + 1$. For $\epsilon = 1/2$ we choose $n = \max(k_1 + 1, n_\epsilon)$ and again by (2.9) we find $k_2 \geq k_1 + 1 > k_1$ with $k_2 \geq n_\epsilon$ such that

$$L - \frac{1}{2} < x_{k_2} < L + \frac{1}{2}.$$

By induction we then find a subsequence $\{x_{k_n}\}$ of $\{x_n\}$ such that $|x_{k_n} - L| < \frac{1}{n} \forall n \geq 1$, hence converging to L . This proves (ii).

Conversely, suppose that (i) and (ii) hold, and let $\epsilon > 0$. From (i) we infer that

$$\sup_{k \geq n} \{x_k\} \leq L + \epsilon \quad \forall n \geq \bar{n}$$

and, since there is a subsequence that converges to L ,

$$\sup_{k \geq n} \{x_k\} \geq L - \epsilon \quad \text{for all } k \text{ large enough.}$$

In conclusion there is n_ϵ such that

$$\left| \sup_{k \geq n} \{x_k\} - L \right| \leq \epsilon \quad \text{for } n \geq n_\epsilon,$$

that is, ϵ being arbitrary, L is the upper limit of $\{x_n\}$. \square

Similarly we have

2.40 Proposition. *Let $\{x_n\}$ be a sequence of real numbers. The number $L \in \mathbb{R}$ is the lower limit of $\{x_n\}$ if and only if*

- (i) $\forall \epsilon > 0 \exists \bar{n}$ such that $x_n > L - \epsilon$ for $n \geq \bar{n}$,
- (ii) *there exists a subsequence $\{x_{k_n}\}$ of $\{x_n\}$ that converges to L .*

Proof. We can give a direct proof following the scheme of the proof of Proposition 2.39 or derived from Proposition 2.39, since

$$\liminf_{n \rightarrow +\infty} x_n = -\limsup_{n \rightarrow +\infty} (-x_n).$$

\square

2.41 ¶. Show the following

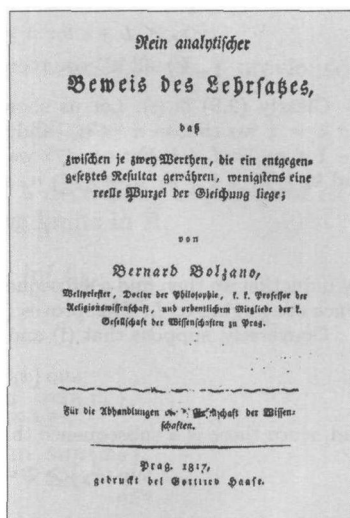
Proposition. *Let $\{x_n\}$ be a sequence in \mathbb{R} .*

- (i) $+\infty$ is the upper limit of $\{x_n\}$ if and only if $\{x_n\}$ has a subsequence that diverges to $+\infty$.
- (ii) $-\infty$ is the upper limit of $\{x_n\}$ if and only if $\{x_n\}$ diverges to $-\infty$.
- (iii) $-\infty$ is the lower limit of $\{x_n\}$ if and only if $\{x_n\}$ has a subsequence that diverges to $-\infty$.
- (iv) $+\infty$ is the lower limit of $\{x_n\}$ if and only if $\{x_n\}$ diverges to $+\infty$.

2.42 ¶. The *limit values* of a sequence $\{x_n\}$ are the limits of the convergent subsequences of $\{x_n\}$. Show that the upper limit (respectively the lower limit) is the supremum (respectively the infimum) of the set of the limit values.



Figure 2.7. Bernhard Bolzano (1781–1848) and the frontispiece of the work where Bolzano–Weierstrass theorem appears.



d. Bolzano–Weierstrass theorem

Since every bounded sequence has a finite upper limit L and a subsequence converging to L , see Proposition 2.39, we can state

2.43 Theorem (Bolzano–Weierstrass). *Every bounded sequence of reals contains a convergent subsequence.*

e. The continuity property of the reals

We have seen that the continuity of reals implies

- (i) the Archimedean property,
- (ii) existence of the limit of monotone sequences,
- (iii) Cantor's principle,
- (iv) Cauchy's criterion,
- (v) existence of upper and lower limits,
- (vi) that every bounded sequence has a convergent subsequence.

Actually we also have

2.44 Proposition. *Let X be an ordered field. The following claims (i) and (ii) are equivalent.*

- (i) *The continuity axiom (C),*
- (ii) *The Archimedean principle and one among*
 - a) *Cantor's principle,*
 - b) *Cauchy's criterion,*
 - c) *existence of limit of monotone sequences,*

- d) existence of the upper and lower limit of a sequence,
 e) every bounded sequence has a convergent subsequence.

2.45 ¶¶. Show Proposition 2.44. [Hint: In order to show that every nonempty bounded set A has a supremum if the Archimedean property and one of the (a), (b), (c), (d) or (e) hold, it is convenient to take into account the following construction. Let $A \subset \mathbb{R}$ be nonempty and bounded. Choose $a_0 \in A$, an upper bound b_0 of A , and set $c = (a_0 + b_0)/2$. If c is an upper bound of A , we set $a_1 := a_0$, $b_1 := c$; otherwise we can find $d \in A$ with $d > c \geq a_0$ and, in this case, we set $a_1 := d$ and $b_1 = b_0$. If we repeat the argument starting from a_1 and b_1 instead of a_0, b_0 , and continue this way, we construct two sequences $\{a_n\}$ and $\{b_n\}$ such that for all n , $a_n \in A$, b_n is an upper bound of A , $a_n \leq b_n$, a_n is increasing, b_n is decreasing, and

$$b_n - a_n \leq (b_0 - a_0)/2^n.$$

Using this construction and one of the (a), (b), (c), (d) or (e) it is not difficult to show the existence of the supremum of A : one needs the Archimedean property to show that the limits of $\{a_n\}$ and $\{b_n\}$ are equal.]

2.3 Limits of Sequences and Continuity

a. Limits of sequences and limits of functions

The definition of limit of a sequence in Section 2.2 and of limit of a function can be reduced one to the other.

2.46 Theorem. Let $f :]a, b[\rightarrow \mathbb{R}$ be a function and $x_0 \in [a, b]$. The following two claims are equivalent:

- (i) $f(x) \rightarrow L \in \overline{\mathbb{R}}$ as $x \rightarrow x_0$, $x \in]a, b[$,
 (ii) for any sequence $\{x_n\} \subset]a, b[\setminus \{x_0\}$ with $x_n \rightarrow x_0$ we have $f(x_n) \rightarrow L$.

Proof. We prove the theorem in the case $L \in \mathbb{R}$ and leave the proof to the reader in the other cases.

(i) \Rightarrow (ii) Let $\epsilon > 0$. By assumption

$$\exists \delta > 0: \text{ if } x \in]a, b[, x \neq x_0 \text{ and } |x - x_0| < \delta, \text{ then } |f(x) - L| < \epsilon. \quad (2.10)$$

If $\{x_n\}$ converges toward x_0 , then there is an index ν such that $|x_n - x_0| < \delta$ for all $n \geq \nu$; since $x_n \neq x_0$, and $x_n \in]a, b[$, (2.10) yields $|f(x_n) - L| < \epsilon$ for all $n \geq \nu$, that is $f(x_n) \rightarrow L$, ϵ being arbitrary.

(ii) \Rightarrow (i) Assume that $f(x)$ has no limit when $x \rightarrow x_0$. Then there exist $\epsilon_0 > 0$ and, for any given $\delta > 0$, a point $x \in]a, b[\setminus \{x_0\}$ such that $|x - x_0| < \delta$ while $|f(x) - L| > \epsilon_0$. Choosing $\delta = 1, 1/2, 1/3, \dots$ we define this way a sequence $\{x_n\}$ with values in $]a, b[\setminus \{x_0\}$ such that

$$|x_n - x_0| < 1/n \quad \text{and} \quad |f(x_n) - L| > \epsilon_0.$$

In particular $x_n \rightarrow x_0$ and $f(x_n)$ does not converge to L : a contradiction. \square

2.47 Example. Consider the sequence $x_n := \sqrt{n} \sin(1/\sqrt{n})$. Since $f(x) = \sin x/x \rightarrow 1$ as $x \rightarrow 0^+$ and $x_n = f(1/\sqrt{n})$, Theorem 2.46 yields $x_n \rightarrow 1$.

2.48 Example. Let $f, g :]a, b[\rightarrow \mathbb{R}$ be two functions, $x_0 \in [a, b]$, and $f(x) \rightarrow L, g(x) \rightarrow M$ as $x \rightarrow x_0, x \in]a, b[$. We may prove that $f(x) + g(x) \rightarrow L + M$ as $x \rightarrow x_0$ as a consequence of Proposition 2.20. For any sequence $\{x_n\} \subset]a, b[\setminus \{x_0\}$, which converges to x_0 , Theorem 2.46 yields $f(x_n) \rightarrow L, g(x_n) \rightarrow M$, hence $f(x_n) + g(x_n) \rightarrow L + M$, according to Proposition 2.20. Again Theorem 2.46 then allows us to conclude that

$$f(x) + g(x) \rightarrow L + M \quad \text{as } x \rightarrow x_0, \quad x \in]a, b[.$$

2.49 Example. Let us give another example proving the change of variable rule, Proposition 2.27 of [GM1].

Proposition. Let $f : I \rightarrow \mathbb{R}$ be a function defined on an interval I , let x_0 be a point in I or one of its extremal points, and let $f(x) \rightarrow L, L \in \overline{\mathbb{R}}$, as $x \rightarrow x_0$. Let $x(t) : J \rightarrow I$ be a function defined in an interval J onto I such that $x(t) \rightarrow x_0$ as $t \rightarrow t_0, t_0$ being a point in J or one of its extremal points. If one of the following two conditions holds:

- (i) $x_0 \in I$ and f is continuous at x_0 ,
- (ii) $x(t)$ never takes the value x_0 for $t \neq t_0$,

then $f(x(t)) \rightarrow L$ as $t \rightarrow t_0, t \in J$.

Proof. Let $\{t_n\}$ be a sequence with values in $J \setminus \{t_0\}$ that converges to t_0 . Clearly $\{x(t_n)\} \subset I$ and, by Theorem 2.46, $x(t_n) \rightarrow x_0$. Let us prove that $f(x(t_n)) \rightarrow L$.

If x never takes the value x_0 for $t \neq t_0$, then $x(t_n) \in I \setminus \{x_0\}$ and therefore $f(x(t_n)) \rightarrow L$ by Theorem 2.46.

If $f(x_0) = L$, for the subsequence $\{s_n\}$ of $\{t_n\}$ such that $f(s_n) \neq x_0$ we have $f(x(s_n)) \rightarrow L$ on account of Theorem 2.46. Therefore for any $\epsilon > 0$ there exists \bar{n} such that $|f(x(x_n)) - L| < \epsilon$ for any $n \geq \bar{n}$ such that $x(t_n) \neq x_0$. Since $f(x_0) = L$, $|f(x(t_n)) - L| < \epsilon$ for all $n \geq \bar{n}$. That is, $f(x(t_n)) \rightarrow L, \epsilon$ being arbitrary.

Finally, since $f(x(t_n)) \rightarrow L$ for any sequence $\{x_n\} \subset J \setminus \{t_0\}$, the claim follows applying once again Theorem 2.46. \square

b. Continuity in terms of sequences

Let $f : [a, b] \rightarrow \mathbb{R}$ be a function and $x_0 \in [a, b]$. We recall, see Chapter 2 of [GM1], that f is *continuous* at x_0 , if $f(x) \rightarrow f(x_0)$ as $x \rightarrow x_0, x \in [a, b]$, or, in the ϵ - δ language

$$\begin{aligned} \forall \epsilon > 0 \exists \delta > 0 : & \text{ if } x \in [a, b] \text{ and } |x - x_0| < \delta, \\ & \text{ then } |f(x) - f(x_0)| < \epsilon. \end{aligned}$$

Theorem 2.46 yields at once

2.50 Proposition. Let $f : [a, b] \rightarrow \mathbb{R}$. f is continuous at $x_0 \in [a, b]$ if and only if for every sequence $\{x_n\} \subset [a, b]$ with $x_n \rightarrow x_0$ we have $f(x_n) \rightarrow f(x_0)$.

In terms of sequences we can also give proofs of the intermediate value theorem and of Weierstrass's theorem that are more robust, i.e., that can be extended to more general contexts.

2.51 Theorem. Let $f : [a, b] \rightarrow \mathbb{R}$ be a continuous function on $[a, b]$. If $f(a) < 0$ and $f(b) > 0$, then there exists $x_0 \in [a, b]$ such that $f(x_0) = 0$.

Proof. Set $a_0 := a$, $b_0 := b$ and $c := (a_0 + b_0)/2$. If $f(c) = 0$, then c is the zero x_0 ; otherwise we set

$$\begin{cases} a_1 := c, & b_1 := b & \text{if } f(c) < 0, \\ a_1 := a, & b_1 := c & \text{if } f(c) > 0. \end{cases}$$

The function f is continuous on $[a_1, b_1] \subset [a_0, b_0]$, $f(a_1) < 0$ and $f(b_1) > 0$. Repeating the argument with a_1, b_1 instead of a_0 and b_0 , and proceeding this way, we either find c with $f(c) = 0$ after a finite number of steps, or we construct two sequences $\{a_n\}$ and $\{b_n\}$ with the following properties:

- (i) $a_0 = a$, $b_0 = b$,
- (ii) a_n is increasing, b_n is decreasing,
- (iii) $f(a_n) < 0$ and $f(b_n) > 0$,
- (iv) $|b_n - a_n| = \frac{1}{2}|b_{n-1} - a_{n-1}| = 2^{-n}|b_0 - a_0|$.

When $n \rightarrow \infty$ $a_n \uparrow x_0$, $b_n \downarrow y_0$, and $|y_0 - x_0| \leq |b_n - a_n| \forall n$. Since by (iv) $|b_n - a_n| \rightarrow 0$, we in fact have $x_0 = y_0$, and f being continuous, $f(a_n) \rightarrow f(x_0)$ and $f(b_n) \rightarrow f(x_0)$. On the other hand, by the constancy of sign, according to (ii) we infer $f(x_0) \leq 0$ and $f(x_0) \geq 0$. We therefore conclude that we must have $f(x_0) = 0$. \square

2.52 Theorem (Weierstrass). Every continuous function $f : [a, b] \rightarrow \mathbb{R}$ on a closed and bounded interval attains its maximum and its minimum value.

Proof. Let us prove that f attains its minimum. Define $E := f([a, b])$ and let $L := \inf E$ and let $\{y_n\}$ be a minimizing sequence for E , that is $\{y_n\} \subset E$ and $y_n \rightarrow L$. Since $E = f([a, b])$, there is also a sequence $\{x_n\} \subset [a, b]$ such that $f(x_n) = y_n \forall n$. The sequence $\{x_n\}$ is clearly bounded and therefore, by the Bolzano–Weierstrass theorem, contains a subsequence $\{x_{k_n}\}$ that converges to some point $x_0 \in \mathbb{R}$. Actually, $[a, b]$ being a closed interval, $x_0 \in [a, b]$.

We shall now prove that x_0 is a minimizer for f . Since $\{x_{k_n}\}$ is a subsequence of $\{x_n\}$, from one side $f(x_{k_n}) - y_{k_n} \rightarrow L$; on the other hand $f(x_{k_n}) \rightarrow f(x_0)$, f being continuous. Uniqueness of the limit yields $f(x_0) = L$, that is the claim. \square

2.4 Some Special Sequences

In this section we discuss a number of sequences that turn out to be quite relevant for the sequel.



Figure 2.8. John Wallis (1616–1703) and the frontispiece of his *Opera Mathematica*.

a. Elementary limits

2.53 Example (Geometric sequence). Let $x_n := q^n$, $n \geq 0$, $q \in \mathbb{R}$. If $q = 1$, then trivially $q^n = 1 \forall n$ and $q^n \rightarrow 1$. If $q = -1$, then $q^n = (-1)^n$ has no limit.

Proposition. We have

- (i) $q^n \rightarrow \infty$ if $q > 1$,
- (ii) $q^n \rightarrow 0$ if $|q| < 1$,
- (iii) q^n has no limit if $q < -1$.

Proof. Let $q > 1$. Since $q^x \rightarrow +\infty$ as $x \rightarrow +\infty$, $x \in \mathbb{R}$ (see, for example, [GM1]), we conclude that $q^n \rightarrow \infty$ by Theorem 2.46.

For a direct proof, which makes no use of calculus, set $q = 1 + h$, $h > 0$, and from Newton's binomial, Example 2.6, or from Bernoulli's inequality, Example 1.34, we have

$$q^n = (1 + h)^n \geq nh,$$

thus $q^n \rightarrow \infty$ by the comparison test (see, for example, Proposition 2.18).

If $|q| < 1$, we write $|q| = 1/(1 + h)$ with $h > 0$ to find

$$|q|^n \leq \frac{1}{(1 + h)^n} \leq \frac{1}{nh} \rightarrow 0$$

thus $q^n \rightarrow 0$, again by the comparison test. Finally if $q < -1$, the sequence has no limit since its two subsequences q^{2n} and q^{2n+1} have different limits

$$q^{2n} = |q|^{2n} \rightarrow +\infty, \quad q^{2n+1} = -|q|^{2n+1} \rightarrow -\infty.$$

□

2.54 Example. $\sqrt[n]{n} \rightarrow 1$. In fact, for $x \in \mathbb{R}$, $x > 0$, we can write $x^{1/x} = \exp(\log x/x)$. Since the exponential function is continuous and $\log x/x \rightarrow 0$ as $x \rightarrow +\infty$ (see, for example, [GM1]), we then conclude that

$$x^{1/x} \rightarrow 1 \quad \text{as } x \rightarrow +\infty.$$

Consequently $\sqrt[n]{n} \rightarrow 1$.

We may also proceed as follows. We observe that $\sqrt[n]{n} \geq 1$, therefore $\sqrt[n]{n} =: 1 + h_n$, with $h_n \geq 0$, that is

$$(1 + h_n)^n = n.$$

Using the binomial theorem, Example 2.6, we deduce

$$n = (1 + h_n)^n = 1 + \frac{n(n-1)}{2} h_n^2 + \text{positive terms} \geq 1 + \frac{n(n-1)}{2} h_n^2,$$

which yields

$$0 \leq h_n^2 \leq \frac{2}{n}$$

or

$$0 \leq h_n = \sqrt[n]{n} - 1 \leq \sqrt{\frac{2}{n}},$$

from which we get $\sqrt[n]{n} \rightarrow 1$ since $1/\sqrt{n} \rightarrow 0$.

2.55 Example. $\sqrt[n]{a} \rightarrow 1$ for all $a > 0$. If $a > 1$, we have

$$1 \leq \sqrt[n]{a} \leq \sqrt[n]{n} \quad \text{for } n \geq a.$$

Since $\sqrt[n]{n} \rightarrow 1$, the squeezing test yields $\sqrt[n]{a} \rightarrow 1$. If $0 < a < 1$ we write

$$\sqrt[n]{a} = \frac{1}{\sqrt[n]{\frac{1}{a}}}$$

to get

$$\lim_{n \rightarrow \infty} \sqrt[n]{a} = \frac{1}{\lim_{n \rightarrow \infty} \sqrt[n]{\frac{1}{a}}} = \frac{1}{1} = 1.$$

The claims in Examples 2.56 and 2.57 below are part of a series of results known as *Cesàro theorems*.

2.56 Example. We have

Proposition. If $a_n \rightarrow L$, then $\frac{1}{n} \sum_{j=1}^n a_j \rightarrow L$.

Proof. Assume $L \in \mathbb{R}$ and proceed similarly in the other cases. Given $\epsilon > 0$, we find \bar{n} such that $|a_i - L| < \epsilon$ for all $i \geq \bar{n}$. On the other hand

$$\left| \frac{1}{n} \sum_{i=1}^n a_i - L \right| = \left| \frac{1}{n} \sum_{i=1}^n (a_i - L) \right| \leq \frac{1}{n} \sum_{i=1}^n |a_i - L|,$$

hence for $n > \bar{n}$,

$$\left| \frac{1}{n} \sum_{i=1}^n a_i - L \right| \leq \frac{1}{n} \sum_{i=1}^{\bar{n}-1} |a_i - L| + \frac{1}{n} \sum_{i=\bar{n}}^n |a_i - L| \leq \frac{1}{n} \sum_{i=1}^{\bar{n}-1} |a_i - L| + \frac{(n - \bar{n} + 1)\epsilon}{n};$$

thus we conclude that $\left| \frac{1}{n} \sum_{i=1}^n a_i - L \right| < 2\epsilon$ for n sufficiently large. \square

Notice that $\left\{ \frac{1}{n} \sum_{i=1}^n a_i \right\}$ may have a limit, while $\{a_n\}$ has no limit, as the sequence $\{(-1)^n\}$ shows.

2.57 Example. We have

Proposition. Let $\{a_n\}$ be a sequence of positive numbers and let $L \in \overline{\mathbb{R}}$. If

$$\frac{a_{n+1}}{a_n} \longrightarrow L,$$

then $\sqrt[n]{a_n} \rightarrow L$.

Proof. Suppose $0 < L < \infty$; the cases $L = 0$ and $L = +\infty$ are simpler, and we leave them to the reader. Set for $n \geq 0$,

$$b_n := \log \frac{a_{n+1}}{a_n}$$

and observe that $b_n \rightarrow \log L$ and that

$$\begin{aligned} \frac{1}{n} \log(a_n) &= \frac{1}{n} \left(\log a_0 + (\log a_1 - \log a_0) + (\log a_2 - \log a_1) + \cdots \right. \\ &\quad \left. + (\log a_n - \log a_{n-1}) \right) \\ &= \frac{1}{n} \log a_0 + \frac{1}{n} \sum_{i=1}^{n-1} b_i. \end{aligned}$$

Therefore $\frac{1}{n} \log(a_n) \rightarrow 0 + \log L$, on account of Example 2.56. This proves the claim.

Let us give a more direct proof which makes no use of logarithms. Let $0 < \epsilon < L$. Then we can find \bar{n} such that for all $n \geq \bar{n}$,

$$(L - \epsilon) a_n < a_{n+1} < (L + \epsilon) a_n,$$

and, by iteration,

$$(L - \epsilon)^n \frac{a_{\bar{n}}}{(L - \epsilon)^{\bar{n}}} < a_n < (L + \epsilon)^n \frac{a_{\bar{n}}}{(L + \epsilon)^{\bar{n}}},$$

that is,

$$(L - \epsilon) \sqrt[n]{B} < \sqrt[n]{a_n} < (L + \epsilon) \sqrt[n]{C} \quad \text{for } n \geq \bar{n}$$

where

$$B := \frac{a_{\bar{n}}}{(L - \epsilon)^{\bar{n}}} \quad \text{and} \quad C := \frac{a_{\bar{n}}}{(L + \epsilon)^{\bar{n}}}$$

depend on ϵ , but not on n . Since $\sqrt[n]{B}, \sqrt[n]{C} \rightarrow 1$, there is \bar{n}_1 such that

$$L - 2\epsilon < (L - \epsilon) \sqrt[n]{B}, \quad \text{and} \quad (L + \epsilon) \sqrt[n]{C} < L + 2\epsilon$$

for $n \geq \bar{n}_1$. We therefore conclude for $n \geq \max(\bar{n}, \bar{n}_1)$,

$$L - 2\epsilon < \sqrt[n]{a_n} < L + 2\epsilon,$$

which yields the claim, since ϵ is arbitrary. \square

2.58 Example. Let x_n be the quotient of two polynomials at n ,

$$x_n = \frac{a_p n^p + a_{p-1} n^{p-1} + \cdots + a_1 n + a_0}{b_q n^q + b_{q-1} n^{q-1} + \cdots + b_1 n + b_0},$$

p and q being the degrees, that is, $a_p, b_q \neq 0$. Of course numerator and denominator diverge, but by factorizing n^p in the numerator and n^q in the denominator we find

$$x_n = n^{p-q} \frac{a_p + a_{p-1} \frac{1}{n} + \cdots + a_1 \frac{1}{n^{p-1}} + a_0 \frac{1}{n^p}}{b_q + b_{q-1} \frac{1}{n} + \cdots + b_1 \frac{1}{n^{q-1}} + b_0 \frac{1}{n^q}}.$$

The second factor, the largest fraction on the right, tends to a_p/b_q . We then conclude that

$$x_n \rightarrow \begin{cases} +\infty \cdot \operatorname{sgn}(a_p/b_q) & \text{if } p > q, \\ a_p/b_q & \text{if } p = q, \\ 0 & \text{if } p < q. \end{cases}$$

b. Powers, exponentials and factorials

2.59 Example (Exponentials grow faster than powers). Let $x_n := n^k/q^n$ with $k \in \mathbb{N}$ and $q > 1$. We claim that

$$\frac{n^k}{q^n} \rightarrow 0.$$

For that, write $q = 1 + h$ with $h > 0$ and apply the binomial theorem with $n \geq k + 1$ to find

$$q^n = (1 + h)^n \geq \binom{n}{k+1} h^{k+1} + \text{positive terms} \geq \frac{n(n-1)(n-2) \cdots (n-k)}{(k+1)!},$$

that is

$$0 \leq \frac{n^k}{q^n} \leq \frac{(k+1)! n^k}{h^{k+1} n(n-1)(n-2) \cdots (n-k)}.$$

This yields the claim by comparison, since the right-hand side is the quotient of two polynomials respectively of degree k and $k + 1$, and thus tends to zero, according to Example 2.58.

2.60 Example (Factorials grow faster than powers). $x_n := \frac{n^k}{n!} \rightarrow 0$. For that observe that $n! = n(n-1)(n-2)(n-3) \cdots 3 \cdot 2 \cdot 1 \geq n(n-1)(n-2)(n-k)$ if $n \geq k + 1$. The comparison test and Example 2.59 then yield

$$0 \leq \frac{n^k}{n!} \leq \frac{n^k}{n(n-1) \cdots (n-k)} \rightarrow 0.$$

2.61 Example. $x_n := n!/n^n \rightarrow 0$. It suffices to observe that

$$\frac{n!}{n^n} = \frac{n}{n} \frac{n-1}{n} \cdots \frac{1}{n} \leq \frac{1}{n}.$$

2.62 Example (Factorials grow faster than exponentials). $x_n := \frac{q^n}{n!} \rightarrow 0$. This is trivial if $0 < q \leq 1$. If $q > 1$, we observe that

$$\frac{q^n}{n!} = \frac{q}{n} \frac{q}{n-1} \cdots \frac{q}{1}$$

and that all factors q/n with $n > q$ are smaller than 1. Thus

$$0 \leq \frac{q^n}{n!} = \frac{q}{n} \frac{q}{[q]} \frac{q}{[q]-1} \cdots \frac{q}{1} =: c(q) \frac{q}{n} \quad \text{for } n \geq q,$$

where $[q]$ denotes the integer part of q , and this yields the result trivially.

2.63 Example (Euler's number). For any $x \in \mathbb{R}$ we have

$$\lim_{n \rightarrow \infty} \left(1 + \frac{x}{n}\right)^n = e^x. \quad (2.11)$$

The claim is trivial if $x = 0$. For $x \neq 0$, recall that $De^x = e^x$ (see, for example, Section 4.3 of [GM1]), a claim which is equivalent to

$$\lim_{x \rightarrow 0} \frac{e^x - 1}{x} = 1, \quad \text{or to} \quad \lim_{x \rightarrow 0} \frac{\log(1+x)}{x} = 1, \quad (2.12)$$

compare 4.3 of [GM1]. By the change of variables $y = 1/t$, $t > 0$, this yields

$$\lim_{t \rightarrow +\infty} t \log \left(1 + \frac{1}{t}\right) = 1$$

or

$$\lim_{t \rightarrow \infty} \left(1 + \frac{1}{t}\right)^t = e,$$

consequently

$$\lim_{t \rightarrow \infty} \left(1 + \frac{x}{t}\right)^t = \left(\lim_{t \rightarrow \infty} \left(1 + \frac{x}{t}\right)^{\frac{t}{x}}\right)^x = e^x.$$

Thus Theorem 2.46 yields (2.11).

From (2.11) we get in particular

$$e = \lim_{n \rightarrow \infty} \left(1 + \frac{1}{n}\right)^n \quad (2.13)$$

which in fact is trivially equivalent to (2.11). In Section 2.5 below we will deduce (2.12), and consequently (2.11), directly from (2.13) making no use of calculus.

2.64 Example (Compound interest). If d is the rate of interest per cent and interests are capitalized every year, the accumulated amount of an original capital x after n years is given by

$$\begin{cases} x_0 = x, \\ x_{n+1} = x_n(1 + d), \end{cases}$$

which yields

$$x_n = x(1 + d)^n.$$

If the interests are capitalized N times per year, then we have

$$x_n = x \left(1 + \frac{d}{N}\right)^{Nn}$$

and for a continuous capitalization,

$$x_n := x \lim_{N \rightarrow \infty} \left(1 + \frac{d}{N}\right)^{Nn} = xe^{nd},$$

according to Example 2.63.

2.65 Example (Sum of the terms of a geometric progression). Let $q \in \mathbb{R}$. Let us compute the sum

$$G_q(n) := \sum_{j=0}^n q^j = 1 + q + q^2 + \cdots + q^n.$$

If $q = 1$, then $G_1(n) = n + 1$. For $q \neq 1$ the following formula holds:

$$G_q(n) := \sum_{j=0}^n q^j = \frac{1 - q^{n+1}}{1 - q}, \quad \forall n \geq 0 \quad (2.14)$$

as it is easily seen multiplying both sides by $1 - q$,

$$\begin{aligned} (1 - q)G_q(n) &= (1 - q) \sum_{j=0}^n q^j = \sum_{j=0}^n q^j - q \sum_{j=0}^n q^j \\ &= 1 + q + q^2 + \cdots + q^n - q - q^2 - q^3 - \cdots - q^{n+1} = 1 - q^{n+1}. \end{aligned}$$

Formula (2.14) yields

$$G_q(n) \rightarrow \begin{cases} +\infty & \text{if } q \geq 1, \\ \frac{1}{1 - q} & \text{if } |q| < 1, \\ \text{does not exist} & \text{if } q \leq -1. \end{cases} \quad (2.15)$$

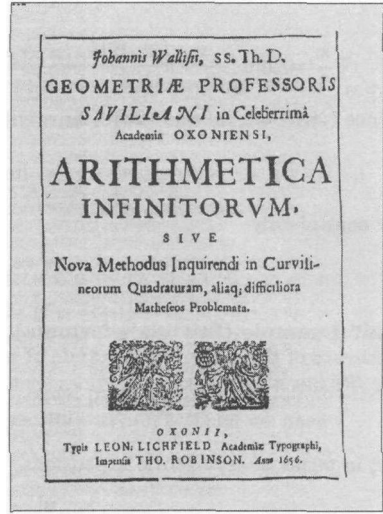


Figure 2.9. The frontispiece of *Arithmetica infinitorum* by John Wallis (1616–1703).

c. Wallis and Stirling formulas

2.66 Example (Wallis's formula). For $n = 1, 2, \dots$, define J_n by

$$J_n := \int_0^{\pi/2} \sin^n x \, dx.$$

Since an integral of $\sin^n x$, $I_n(x)$, is given (see, for example, Example 4.34 of [GM1]) by

$$\begin{cases} I_0(x) = x, & I_1(x) = -\cos x, \\ I_n(x) = \frac{n-1}{n} I_{n-2}(x) - \frac{\sin^{n-1} x \cos x}{n} \quad \forall n \geq 2, \end{cases}$$

we have $J_n = I_n(\pi/2) - I_n(0)$, i.e.,

$$J_0 = \frac{\pi}{2}, \quad J_1 = 1, \quad J_n = \frac{n-1}{n} J_{n-2} \quad \forall n \geq 2. \quad (2.16)$$

Consequently

$$J_{2n} = \frac{2n-1}{2n} \frac{2n-3}{2n-2} \cdots \frac{1}{2} \cdot \frac{\pi}{2}, \quad \text{and} \quad J_{2n+1} = \frac{2n}{2n+1} \frac{2n-2}{2n-1} \cdots \frac{2}{3} \cdot 1,$$

which we can write, introducing the *semifactorials*

$$(2n)!! := 2n(2n-2) \cdots 4 \cdot 2, \quad (2n+1)!! := (2n+1)(2n-1) \cdots 5 \cdot 3 \cdot 1,$$

as

$$J_{2n} = \frac{(2n-1)!!}{(2n)!!} \frac{\pi}{2}, \quad J_{2n+1} = \frac{(2n)!!}{(2n+1)!!}.$$

Since $\sin^{k+1} x \leq \sin^k x$ for all $k \geq 1$, we have $J_{k+1} \leq J_k \quad \forall k$, hence

$$1 \leq \frac{J_{2n}}{J_{2n+1}} = \frac{2n+1}{2n} \frac{J_{2n}}{J_{2n-1}} \leq 1 + \frac{1}{2n}.$$

Consequently $J_{2n}/J_{2n+1} \rightarrow 1$: this is *Wallis's formula* for π :

$$\frac{\pi}{2} = \lim_{n \rightarrow \infty} \frac{[(2n)!!]^2}{(2n+1)!!(2n-1)!!} = \lim_{n \rightarrow \infty} \frac{2 \ 2 \ 4 \ 4 \ \dots \ (2n)}{1 \ 3 \ 3 \ 5 \ \dots \ (2n-1)} \frac{(2n)}{(2n+1)}. \quad (2.17)$$

Since $(2n)!! = 2^n n!$ and $(2n)!!(2n-1)!! = (2n)!$, we can rewrite (2.17) as

$$\frac{\pi}{2} = \lim_{n \rightarrow \infty} \frac{2^{4n} (n!)^4}{(2n)! (2n+1)}$$

or equivalently

$$\sqrt{\pi} = \lim_{n \rightarrow \infty} \frac{2^{2n} (n!)^2}{(2n)! \sqrt{n}}. \quad (2.18)$$

2.67 Example (Stirling's formula). In many applications, notably in statistics, an estimate of the order of magnitude of $n!$ for n large is useful. This estimate is provided by *Stirling's formula*

$$\lim_{n \rightarrow \infty} \frac{n!}{n^n e^{-n} \sqrt{n}} = \sqrt{2\pi}$$

or, in terms of asymptotic expansions,

$$n! \sim n^n e^{-n} \sqrt{2\pi n}.$$

In order to prove this, set

$$a_n := \frac{n!}{n^n e^{-n} \sqrt{n}}, \quad n \geq 1,$$

and observe that the real function

$$f(t) := \frac{2+t}{2t} \log(1+t), \quad 0 < t < 1,$$

is strictly increasing, $f(t) \rightarrow 1$ as $t \rightarrow 0^+$ and $f(t) \leq 1 + t^2/4$ since $\log(1+t) \leq t - t^2/2 + t^3/3$. Therefore

$$1 \leq \frac{a_n}{a_{n+1}} = \frac{1}{e} \left(1 + \frac{1}{n}\right)^{n+1/2} = \frac{1}{e} \exp(f(1/n)) \leq e^{1/(4n^2)}. \quad (2.19)$$

From (2.19) we see that $\{a_n\}$ is strictly decreasing, thus it converges to a limit L , $L \geq 0$, and that $L > 0$ since $a_n e^{-1/n}$ is strictly increasing to L . In order to compute L , we observe that

$$\frac{a_n^2}{a_{2n} \sqrt{2}} = \frac{(n!)^2 2^{2n}}{(2n)! \sqrt{n}}.$$

Passing to the limit and taking into account Wallis's formula (2.18), we find

$$\frac{L^2}{L\sqrt{2}} = \sqrt{\pi} \quad \text{i.e.,} \quad L = \sqrt{2\pi}.$$

Notice that the limits

$$\frac{n!}{n^n a^n} \rightarrow \begin{cases} 0 & \text{if } a > e, \\ +\infty & \text{if } a < e \end{cases}$$

are easy to obtain, instead. In fact, if $a_n := \frac{n!}{n^n a^n}$, we have

$$\frac{a_{n+1}}{a_n} = \frac{(n+1)n!}{\frac{(n+1)^{n+1} a^{n+1}}{n!}} = \frac{1}{a} \left(1 + \frac{1}{n}\right)^n \rightarrow \frac{e}{a}.$$

This implies that $\{a_n\}$ grows or decays exponentially fast if $a < e$ or $a > e$ respectively.

d. Numerical integration

Let $f : [a, b] \rightarrow \mathbb{R}$ be a Riemann integrable function, and $a = x_0 < x_1 < \dots < x_n = b$ be a subdivision of $[a, b]$. Then the integral of f , see, e.g., Section 3.2.1 of [GM1], is the limit of the sums

$$\sum_{i=0}^{n-1} (x_{i+1} - x_i) f(\xi_i)$$

when the length $\sup_i |x_{i+1} - x_i|$ of the partition $a = x_0 < x_1 < \dots < x_n = b$ tends to zero, ξ being any point chosen in $[x_i, x_{i+1}]$.

This allows us to find approximate formulas for the integral of f .

2.68 The rectangle rule. Divide $[a, b]$ into N equal parts by means of the subdivision $x_i = a + i \frac{b-a}{N}$, $i = 0, \dots, N$, and choose $\xi_i := (x_i + x_{i+1})/2$. Then we have

$$\int_a^b f(x) dx = \frac{b-a}{N} \sum_{i=0}^{N-1} f(\xi_i) + O(1) \quad \text{as } N \rightarrow \infty.$$

Assuming that f is regular, it is possible to estimate the error

$$E(h) := \int_a^b f(x) dx - \frac{b-a}{N} \sum_{i=0}^{N-1} f(\xi_i)$$

in terms of $h := (b-a)/N$. Suppose for instance that $f \in C^1([a, b])$ then

$$\begin{aligned} |E(h)| &\leq \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} \left| f(x) - f\left(\frac{x_i + x_{i+1}}{2}\right) \right| dx \\ &\leq \sum_{i=0}^{N-1} \sup_{|x_i, x_{i+1}|} |f'| \int_{x_i}^{x_{i+1}} \left| x - \frac{x_i + x_{i+1}}{2} \right| dx \\ &= \frac{1}{2} \sum_{i=0}^{N-1} \sup_{|x_i, x_{i+1}|} |f'| (x_{i+1} - x_i)^2 \\ &= \sup_{|a, b|} |f'| \frac{(b-a)^2}{2N} = (b-a) \sup_{|a, b|} |f'| \frac{h}{2}. \end{aligned}$$

2.69 The trapezoid formula. Here we choose as approximate value of the integral the integral of the piecewise interpolate \hat{f} of f at the points $x_i := a + i \frac{b-a}{N}$, that is of

$$\hat{f}(x) := f(x_i) + \frac{f(x_{i+1}) - f(x_i)}{h} (x - x_i), \quad x_i := a + hi, \quad x_i \leq x \leq x_{i+1}.$$

Therefore

$$\int_a^b \hat{f} dx = \frac{(b-a)}{N} \sum_{i=0}^{N-1} \frac{f(x_i) + f(x_{i+1})}{2}.$$

If f is of class C^2 , by integrating by parts twice and using that $\hat{f}(x_i) = f(x_i) \forall i$, we find

$$\int_{x_i}^{x_{i+1}} (f - \hat{f}) dx = \frac{1}{2} \int_{x_i}^{x_{i+1}} (x - x_i)(x - x_{i+1}) f''(x) dx,$$

hence, being as above $h := (b-a)/N$,

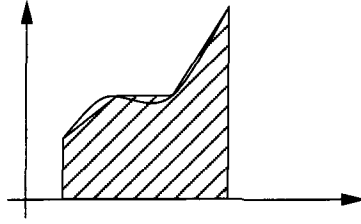


Figure 2.10.

$$\begin{aligned} \left| \int_{x_i}^{x_{i+1}} (f - \hat{f}) dx \right| &\leq \frac{1}{2} \sup_{]x_i, x_{i+1}[} |f''| \int_{x_i}^{x_{i+1}} (x - x_i)(x_{i+1} - x) dx \\ &= \frac{1}{12} \sup_{]x_i, x_{i+1}[} |f''| (x_{i+1} - x_i)^3 = \frac{1}{12} \sup_{]x_i, x_{i+1}[} |f''| h^3. \end{aligned}$$

In conclusion we have

$$\left| \int_a^b f dx - \int_a^b \hat{f} dx \right| \leq \sup_{]a, b[} |f''| \frac{(b-a)^3}{12N^2}$$

or

$$\int_a^b f dx = \frac{b-a}{N} \sum_{i=0}^{N-1} \frac{f(x_i) + f(x_{i+1})}{2} + O\left(\frac{1}{N^2}\right).$$

2.70 Simpson's rule. Instead of interpolating with a piecewise linear function, we want to interpolate with a quadratic function. Let $f: [-1, 1] \rightarrow \mathbb{R}$; we look for $\hat{f}(t) = At^2 + Bt + C$, $t \in [-1, 1]$, in such a way that $\hat{f}(-1) = f(-1)$, $\hat{f}(0) = f(0)$, $\hat{f}(1) = f(1)$. We easily find

$$A = \frac{f(-1) + f(1)}{2} - f(0), \quad B = \frac{f(1) - f(-1)}{2}, \quad C = f(0),$$

consequently

$$\int_{-1}^1 \hat{f} dt = \frac{1}{3} (f(-1) + 4f(0) + f(1)).$$

Set now $x_k := a + kh$, $h := \frac{b-a}{N}$, $N = 2n + 1$ odd, $k = 0, \dots, 2n$. From the above we then infer (by changing variable)

$$\begin{aligned} \int_{x_0}^{x_2} \hat{f} dx &= \frac{h}{3} (f(x_0) + 4f(x_1) + f(x_2)), \\ \int_{x_2}^{x_4} \hat{f} dx &= \frac{h}{3} (f(x_2) + 4f(x_3) + f(x_4)), \\ &\dots \\ \int_{x_{2n-2}}^{x_{2n}} \hat{f} dx &= \frac{h}{3} (f(x_{2n-2}) + 4f(x_{2n-1}) + f(x_{2n})), \end{aligned}$$

that is

$$\int_a^b \hat{f} dx = \frac{h}{3} \left(f(a) + 2 \sum_{k=1}^{n-1} f(2k) + 4 \sum_{k=0}^{n-1} f(2k+1) + f(b) \right).$$

In conclusion we find

$$\int_a^b f \, dx = \int_a^b \hat{f} \, dx + E(h)$$

known as *Simpson's rule*. One can show that, if $f \in C^4([a, b])$, then

$$|E(h)| \leq \sup_{]a, b[} |f^{(4)}| \frac{(b-a)^5}{180N^4}, \quad (2.20)$$

in particular $E(h) = O(h^4)$ as $h \rightarrow 0$.

2.71 ¶¶. Prove (2.20). [*Hint:* Set

$$R(h) = \int_{\bar{x}-h}^{\bar{x}+h} \left(f(x) - \frac{h}{3}(f(\bar{x}-h) + 4f(\bar{x}) + f(\bar{x}+h)) \right) dx.$$

Show that $R'''(h) = -\frac{2}{3}h^2 f^{(4)}(\xi)$ for some $\xi \in]x-h, x+h[$. From $R(0) = R'(0) = R''(0) = 0$ infer then that $R(h) = -\frac{h^5}{90} f^{(4)}(\xi)$ for some $\xi \in [x-h, x+h]$, and hence the result.]

2.5 An Alternative Definition of Exponentials and Logarithms

In [GM1] we defined Euler's number e , the exponential and logarithm functions by making use of the differential and integral calculus. For the sake of completeness we give in this section a more direct, though slightly more involved, definition, which makes use of the calculus of limits and of continuous functions. The procedure is as follows: we define a^x for x rational, we then extend "by continuity" $a^x : \mathbb{Q} \rightarrow \mathbb{R}$ to a continuous function from $\mathbb{R} \rightarrow \mathbb{R}$. Then we define the logarithm with base a as the inverse of a^x , which turns out to be continuous because of 2.48 of [GM1].

a. A definition of a^x using continuity

2.72 Rational powers. Let $a > 1$. Taking into account the existence of the n -th root of a real number, we define a^r when r is rational by

$$a^r := \sqrt[q]{a^p}, \text{ if } r = p/q \in \mathbb{Q}; \quad (2.21)$$

in fact, it is not difficult to show that the result of the operation $\sqrt[q]{a^p}$ depends only on the quotient p/q and not on p and q (if $\frac{p}{q} = \frac{p'}{q'}$, then $\sqrt[q]{a^p} = \sqrt[q']{a^{p'}}$). Thus $a^{p/q} = \sqrt[q]{a^p} = (\sqrt[q]{a})^p$. When $a < 1$, we set, still for $r \in \mathbb{Q}$,

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

and, if $a = 1$, we set $a^r = 1^r = 1 \, \forall r \in \mathbb{Q}$. This way a^r is defined for $a > 0$ and $r \in \mathbb{Q}$.

The computational rules for the exponentials of rational numbers are now an easy consequence: if $x, y \in \mathbb{R}$, $x, y > 0$ and $r, s \in \mathbb{Q}$, we have

$$\begin{aligned} x^r y^r &= (xy)^r, & x^r x^s &= x^{r+s}, & (x^r)^s &= x^{rs}, \\ \text{if } 0 < x < y \text{ and } r > 0, & \text{ then } x^r < y^r, \\ \text{if } 0 < x < 1 \text{ and } r < s, & \text{ then } x^s < x^r, \\ \text{if } x > 1 \text{ and } r < s, & \text{ then } x^r < x^s, \\ \frac{x^r}{y^r} &= \left(\frac{x}{y}\right)^r, & \frac{x^r}{x^s} &= x^{r-s}. \end{aligned} \tag{2.22}$$

2.73 Real powers. The idea is to define a^x when $a \in \mathbb{R}$, $a > 0$ and $x \in \mathbb{R}$ as the limit of a^{x_n} where $\{x_n\}$ is a sequence of rational numbers converging to x . For that we need, and in fact it suffices to show that:

- (i) the sequence a^{x_n} converges if $x_n \rightarrow x$,
- (ii) the limit of a^{x_n} does not depend on $\{x_n\}$ itself but only on the limit x .

Finally, we need to show that the new function a^x agrees with the old one if x is rational.

Suppose $a > 1$. Choosing $b := a^{1/n} - 1$ in Bernoulli's inequality,

$$(1 + b)^n \geq 1 + nb, \quad \forall n \in \mathbb{N}, \quad \forall b \geq -1,$$

we deduce

$$a^{1/n} - 1 \leq \frac{a - 1}{n}. \tag{2.23}$$

Also, it is easily seen that

$$|a^r - 1| \leq a^{|r|} - 1, \quad \forall r \in \mathbb{Q}, \tag{2.24}$$

since $a > 1$. The inequalities (2.23) and (2.24) then yield the following.

2.74 Lemma. *If r_n is a sequence of rationals with $r_n \rightarrow 0$, then $a^{r_n} \rightarrow 1$.*

2.75 Proposition. *For any sequence $\{x_n\} \subset \mathbb{Q}$ with $x_n \rightarrow x$, the sequence a^{x_n} has a limit that depends only on x .*

Proof. If $\{x_n\}$ is an increasing sequence that converges to x , then $\{a^{x_n}\}$ is increasing and bounded, therefore has a limit, $a^{x_n} \rightarrow L$. Consider now any sequence of rationals $\{y_n\}$ so that $y_n \rightarrow x$, then

$$|a^{y_n} - L| \leq |a^{y_n} - a^{x_n}| + |a^{x_n} - L| = a^{x_n} |a^{y_n - x_n} - 1| + |a^{x_n} - L|.$$

Since $a^{x_n} \rightarrow L$ and $y_n - x_n \rightarrow 0$, Lemma 2.74 and the comparison test yield $a^{y_n} \rightarrow L$. \square

Because of Proposition 2.75, the function

$$A(x) := \lim_{n \rightarrow \infty} a^{x_n}, \quad \{x_n\} \subset \mathbb{Q}, \quad x_n \rightarrow x, \quad (2.25)$$

is well defined for any x ; moreover $A(x) = a^x$ for any $x \in \mathbb{Q}$: in fact, if we choose $x_n := x \forall n$, then

$$A(x) = (\text{by (2.25)}) = \lim_{n \rightarrow \infty} a^{x_n} = a^x \text{ (by (2.21))}.$$

Therefore $A(x)$ *extends* to the reals the function a^x , $x \in \mathbb{Q}$, and will be denoted again by a^x , $x \in \mathbb{R}$.

Then one defines a^x when $0 < a < 1$ by

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

and $a^x = 1^x := 1$ if $a = 1$.

2.76 Real powers: laws of exponents. If we take into account how the operation of limit behaves with respect to the algebraic operations and the order of \mathbb{R} , it is easy to extend the claims (2.22) and (2.24) to the case in which $r, s \in \mathbb{R}$.

2.77 ¶. Show that (2.22) and (2.24) hold for $r, s \in \mathbb{R}$.

2.78 Continuity of the exponential. The argument in Proposition 2.75 shows also the following.

2.79 Proposition. *If $\{x_n\} \subset \mathbb{R}$, $x_n \rightarrow x$, then $a^{x_n} \rightarrow a^x$.*

Theorem 2.46 then implies that a^x is a continuous function in \mathbb{R} . Since for $a > 1$, $a^n \rightarrow +\infty$ as $n \rightarrow \infty$ and a^x , $x \in \mathbb{R}$, is monotonically increasing, we also infer

$$\inf_{x \in \mathbb{R}} a^x = 0, \quad \sup_{x \in \mathbb{R}} a^x = +\infty, \quad 0 < a \neq 1.$$

This way the *exponential function* is defined for all $a > 0$ and all $x \in \mathbb{R}$ and agrees with the exponential function defined in [GM1], since both are continuous and agree on rationals.

b. Euler's number e

Before proving the differentiability of the exponential function a^x , we need a definition of the Euler number e .

2.80 Proposition. *The sequence $x_n := (1 + 1/n)^n$ is increasing and bounded, $2 \leq x_n < 3$.*

Proof. In fact

$$\begin{aligned} \left(1 + \frac{1}{n}\right)^n &= \sum_{k=0}^n \binom{n}{k} \frac{1}{n^k} = \sum_{k=0}^n \frac{1}{k!} \frac{n(n-1)\cdots(n-k+1)}{n^k} \\ &= \sum_{k=0}^n \frac{1}{k!} \left(1 - \frac{1}{n}\right) \left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right). \end{aligned} \quad (2.26)$$

Since each term in the sum increases with n and moreover, when n increases, new positive terms add to the sum, we infer that $\{x_n\}$ is strictly increasing and $x_n \geq x_1 = 2 \forall n$. Equality (2.26) yields also

$$\left(1 + \frac{1}{n}\right)^n \leq \sum_{k=0}^n \frac{1}{k!}.$$

On the other hand $2^k \leq k!$ for $k \geq 4$, therefore

$$\begin{aligned} \sum_{j=4}^n \frac{1}{j!} &\leq \sum_{j=4}^n 2^{-j} = -1 - 1/2 - 1/4 - 1/8 + G_{1/2}(n) \\ &< -1 - 1/2 - 1/4 - 1/8 + 2 = 1/8, \end{aligned}$$

if we take into account Example 2.65. We then conclude, for all $n \geq 4$,

$$x_n \leq 1 + 1 + \frac{1}{2} + \frac{1}{6} + \sum_{k=4}^n \frac{1}{k!} < 1 + 1 + \frac{1}{2} + \frac{1}{6} + \frac{1}{8} < 2.9.$$

□

The sequence $(1 + 1/n)^n$ has therefore a finite limit, as it is increasing and bounded. Set

$$e := \lim_{n \rightarrow +\infty} \left(1 + \frac{1}{n}\right)^n, \quad (2.27)$$

and notice that

$$2 < e < 2.9.$$

c. Derivative of the exponential

Finally, let us show directly that $Da^x = a^x \log_e a$. First we show that (2.27) yields

2.81 Proposition. *We have*

$$\lim_{x \rightarrow 0} (1 + x)^{1/x} = e. \quad (2.28)$$

Proof. Of course we do not want to differentiate. Therefore we denote by $[x]$ the integral part of x and observe that

$$\left(1 + \frac{1}{[x] + 1}\right)^{[x]} \leq \left(1 + \frac{1}{x}\right)^x \leq \left(1 + \frac{1}{[x]}\right)^{[x]+1} \quad (2.29)$$

and that

$$\left(1 + \frac{1}{n}\right)^{n+1} \rightarrow e, \quad \text{and} \quad \left(1 + \frac{1}{n+1}\right)^n \rightarrow e.$$

Given $\epsilon > 0$ we then find $\bar{n} \in \mathbb{N}$ such that

$$e - \epsilon < \left(1 + \frac{1}{n+1}\right)^n, \quad \text{and} \quad \left(1 + \frac{1}{n}\right)^{n+1} < e + \epsilon \quad \text{for } n \geq \bar{n},$$

and, on account of (2.29),

$$e - \epsilon < \left(1 + \frac{1}{x}\right)^x \leq e + \epsilon \quad \text{for } x \geq \bar{n}.$$

ϵ being arbitrary, this yields

$$\lim_{x \rightarrow +\infty} \left(1 + \frac{1}{x}\right)^x = e. \quad (2.30)$$

On the other hand, if $x < 0$ and $y = -x$,

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

hence

$$\lim_{x \rightarrow -\infty} \left(1 + \frac{1}{x}\right)^x = \lim_{y \rightarrow +\infty} \left(1 + \frac{1}{y-1}\right)^y = e. \quad (2.31)$$

Changing variable $x = 1/y$, from (2.30) and (2.31) we finally infer

$$\lim_{x \rightarrow 0^\pm} (1+x)^{1/x} = e.$$

□

Since $\log x$ is continuous, as it is the inverse of a continuous function, (2.28) can be written as

$$\lim_{x \rightarrow 0} \frac{\log(1+x)}{x} = 1$$

and, again changing variable, $y = e^x - 1$, we get

$$\lim_{x \rightarrow 0} \frac{e^x - 1}{x} = 1,$$

which yields that e^x is differentiable and $De^x = e^x$, and also $Da^x = a^x \log_a x$ (see, for example, 4.3 and 4.4 of [GM1]).

2.6 Summing Up

Limit of sequences

Definition

$x_n \rightarrow L$ as $n \rightarrow \infty$ means: for every neighborhood V of L there is \bar{n} such that $x_n \in V$ for all $n \geq \bar{n}$.

Properties

- UNIQUENESS. The limit is unique if it exists.
- BOUNDEDNESS. If $x_n \rightarrow L \in \mathbb{R}$, then $\{x_n\}$ is bounded.
- SQUEEZING TEST. If $a_n \leq b_n \leq c_n$ and $a_n \rightarrow L$ and $c_n \rightarrow L$, then also $b_n \rightarrow L$.
- COMPARISON TEST. If $a_n \geq b_n \forall n$ and $b_n \rightarrow +\infty$, then also $a_n \rightarrow +\infty$.
- CONSTANCY OF SIGN. If $x_n \rightarrow L$ and $L > 0$, then x_n is positive for n large enough. If $x_n \geq 0$ and $x_n \rightarrow L$, then $L \geq 0$.

Limit of functions and limit of sequences

- Let $f :]a, b[\rightarrow \mathbb{R}$ and $x_0 \in [a, b]$. Then $f(x) \rightarrow L \in \overline{\mathbb{R}}$ as $x \rightarrow x_0$, $x \in]a, b[$, if and only if $f(x_n) \rightarrow L$ for any sequence $\{x_n\} \subset]a, b[\setminus \{x_0\}$ with $x_n \rightarrow x_0$.
- Let $f : [a, b] \rightarrow \mathbb{R}$. f is continuous at $x_0 \in [a, b]$ if and only if $f(x_n) \rightarrow f(x_0)$ for every sequence $\{x_n\} \subset [a, b]$ with $x_n \rightarrow x_0$.

Fundamental theorems

$\{x_n\}$ is a *Cauchy sequence* if $\forall \epsilon \exists \bar{n}$ such that $|x_h - x_k| < \epsilon$ for all $h, k \geq \bar{n}$.

- MONOTONE SEQUENCES Every monotone sequence has limit in $\overline{\mathbb{R}}$.
- MAXIMIZING AND MINIMIZING SEQUENCES For any nonempty $A \subset \mathbb{R}$ there exist a *maximizing sequence*, that is, a sequence $\{x_n\} \subset A$ with $x_n \rightarrow \sup A$, and a *minimizing sequence*, that is, a sequence $\{x_n\} \subset A$ with $x_n \rightarrow \inf A$.
- CAUCHY'S CRITERION $\{x_n\}$ is convergent if and only if $\{x_n\}$ is a Cauchy sequence.
- BOLZANO-WEIERSTRASS THEOREM Every bounded sequence contains a convergent subsequence.

Upper and lower limits

Definition

$$\liminf_{n \rightarrow \infty} x_n := \lim_{n \rightarrow \infty} \inf_{k \geq n} \{x_k\}, \quad \limsup_{n \rightarrow \infty} x_n := \lim_{n \rightarrow \infty} \sup_{k \geq n} \{x_k\},$$

Properties

- $\liminf_{n \rightarrow \infty} x_n, \limsup_{n \rightarrow \infty} x_n$ always exist in $\overline{\mathbb{R}}$, and $\liminf_{n \rightarrow \infty} x_n \leq \limsup_{n \rightarrow \infty} x_n$,
- $x_n \rightarrow L$ if and only if $\liminf_{n \rightarrow \infty} x_n = \limsup_{n \rightarrow \infty} x_n = L$,
- Let $L \in \mathbb{R}$. $L = \limsup_{n \rightarrow \infty} x_n \in \mathbb{R}$ if and only if
 - (i) $\forall \epsilon > 0, \exists \bar{n}$ such that $x_n \leq L + \epsilon$ for all $n \geq \bar{n}$,
 - (ii) there is a subsequence $\{x_{k_n}\}$ of $\{x_n\}$ such that $x_{k_n} \rightarrow L$.
- Let $L \in \mathbb{R}$. $L = \liminf_{n \rightarrow \infty} x_n \in \mathbb{R}$ if and only if
 - (i) $\forall \epsilon > 0, \exists \bar{n}$ such that $x_n \geq L - \epsilon$ for all $n \geq \bar{n}$,
 - (ii) there is a subsequence $\{x_{k_n}\}$ of $\{x_n\}$ such that $x_{k_n} \rightarrow L$.

Some remarkable limits

$$\begin{array}{ll}
q^n \rightarrow \begin{cases} 0 & \text{if } |q| < 1, \\ +\infty & \text{if } q > 1, \end{cases} & \sqrt[n]{n} \rightarrow 1, \\
\sqrt[n]{|a|} \rightarrow 1 \quad \forall a \neq 0, & \frac{n^k}{q^n} \rightarrow 0, \text{ for } k \in \mathbb{N} \text{ e } q > 1, \\
\frac{q^n}{n!} \rightarrow 0 \quad \forall q \geq 0, & \frac{n!}{n^n} \rightarrow 0, \\
\sum_{j=0}^n q^j \rightarrow \begin{cases} +\infty & \text{if } q \geq 1, \\ 1/(1-q) & \text{if } |q| < 1, \\ \text{does not exist} & \text{if } q < -1, \end{cases} & \left(1 + \frac{x}{n}\right)^n \rightarrow e^x \quad \forall x \in \mathbb{R}, \\
(\text{WALLIS}) \frac{[(2n)!!]^2}{(2n+1)!!(2n-1)!!} \rightarrow \frac{\pi}{2}, & (\text{WALLIS}) \frac{2^{2n}(n!)^2}{(2n)!\sqrt{n}} \rightarrow \sqrt{\pi}, \\
(\text{STIRLING}) \frac{n!}{n^n e^{-n} \sqrt{2\pi n}} \rightarrow 1. &
\end{array}$$

Figure 2.11. Some remarkable limits.

2.7 Exercises

2.82 ¶. Let

$$\begin{aligned}
A &:= \{a_n \in \mathbb{R} \mid n \in \mathbb{N}\}, & B &:= \{b_n \in \mathbb{R} \mid n \in \mathbb{N}\}, \\
C &:= \{a_n + b_n \mid n \in \mathbb{N}\}, & D &:= \{a_n b_n \mid n \in \mathbb{N}\}.
\end{aligned}$$

- (i) Show that $\sup C \leq \sup A + \sup B$. Give examples in which the inequality is strict.
(ii) Assume $a_n, b_n \geq 0, \forall n$ and show that $\sup D \leq \sup A \sup B$, observing that the inequality can be strict.

2.83 ¶. Find the infimum and the supremum of some of the following sets:

$$\begin{array}{ll}
\left\{n - \frac{1}{n} \mid n \in \mathbb{N}, n \geq 1\right\}, & \left\{1 - \frac{1}{n} \mid n \in \mathbb{N}, n \geq 1\right\}, \\
\left\{\frac{1}{x} - \frac{1}{y} \mid x, y \in (0, 1)\right\}, & \left\{\left(\frac{1}{2}\right)^n + \left(-\frac{1}{3}\right)^n \mid n \in \mathbb{N}\right\}, \\
\left\{\sin x \mid x \in (0, \pi/2)\right\}, & \left\{\frac{nm}{n^2 + m^2} \mid n, m \in \mathbb{N}, n, m > 0\right\}, \\
\left\{\frac{xy}{x^2 + y^2} \mid x, y \in \mathbb{R} \setminus (0, 0)\right\}, & \left\{n^\alpha e^{-n^2} \mid n \in \mathbb{N}, n \geq 1\right\}, \alpha > 0.
\end{array}$$

2.84 ¶. Show that $\sqrt{2n}, n, \sqrt{n^2 + 2}$ are subsequences of \sqrt{n} , that is, declare the selection function.**2.85 ¶.** Show that

$$\lim_{n \rightarrow \infty} \left(\frac{1}{n} + \frac{1}{\sqrt{n^2 + 1}} + \frac{1}{\sqrt{n^2 + 2}} + \cdots + \frac{1}{\sqrt{n^2 + 2n}} \right) = 2.$$

2.86 ¶. Compute, if they exist, the limits of some of the following sequences:

$$\begin{array}{ll}
 \frac{n - \sin n}{n + \cos n}, & \left(1 + \frac{1}{n!}\right)^n, \\
 \left(1 + \frac{1}{n}\right)^{n!}, & \sqrt{n}(\sqrt{n+1} - \sqrt{n-1}), \\
 \sqrt[n]{3^n + 5^n}, & \frac{\sin(1/n)}{1 - \cos(1/n)}, \\
 \left(\cos \frac{\pi}{n}\right)^n, & \left(\sin \frac{1}{n}\right)^{n+\frac{1}{n}}, \\
 \sin(n) \log(n) \sin(1/n), & \frac{(\cos n + 3)^n}{n^2}, \\
 \sqrt[n]{3 + n + n^2}, & 3^{n+1} - 3\sqrt{1+n^2}, \\
 \sqrt{n^2 - 1} / \sqrt[3]{n^3 + 1}, & \frac{1}{n} \log(n + n^2), \\
 \frac{2^n + 3^n}{n!}, & \frac{n}{\sqrt[3]{8n^3 - n - n}}, \\
 \sqrt{n + \sqrt{n}} - \sqrt{n - \sqrt{n}}, & \left(1 + \frac{2}{n}\right)^n, \\
 \left(1 + \frac{1}{n^2}\right)^n, & \left(1 + \frac{1}{n}\right)^{n^2}, \\
 \left(\frac{n^2 + 6}{n^2}\right)^{n^2}. &
 \end{array}$$

2.87 ¶¶. Compute, if they exist, the limits of the following sequences:

$$\begin{array}{ll}
 \int_n^{n+1} [\pi - 2 \arctan t] n t^2 dt, & \int_0^1 \tan \frac{\sin n^2 x}{n} dx, \\
 \int_n^{n+1} e^{-t^2} dt, & \sqrt{n} \int_0^n e^{-nt^2} \log t dt, \\
 \int_0^1 (\log t)^n dt, & n \int_0^1 x e^{-nx} dx.
 \end{array}$$

2.88 ¶¶. Let $x_n := |1 - 10^{-4}n^2|$. Find the supremum and the infimum of the set $A := \{x_n \mid x_n^2 < 1 - (x_n - 1)^2\}$.

2.89 ¶¶. Compute the upper and the lower limits of the following sequences:

$$\begin{array}{ll}
 a_n := \sqrt[n]{|n - (-2)^{-n}|}, & a_n := \left(1 + \frac{(-1)^n}{n}\right)^n, \\
 a_n := \begin{cases} \sqrt[n]{n^2} & n \text{ odd,} \\ \frac{n-2}{2n-9} & n \text{ even,} \end{cases} & a_n := \begin{cases} \frac{\sqrt{n}-1}{2\sqrt{n+3}} & n \text{ even,} \\ (-1)^{n/2} & n \text{ odd,} \end{cases} \\
 a_n := \text{distance between } n \text{ and the closest square.} &
 \end{array}$$

2.90 ¶. Show that every sequence contains a monotone subsequence.

2.91 ¶¶. Show that $\{x_n\}$ has limit if and only if every subsequence of $\{x_n\}$ contains a further subsequence which has limit and all these limits are equal.

2.92 ¶¶. Show that \mathbb{R}^2 , \mathbb{R}^3 and in general \mathbb{R}^n are complete metric spaces. [Hint: Show that a sequence converges iff the sequences of its components converge.]

2.93 ¶. Discuss the following claims.

- (i) If $\{a_n\}$ converges and $\{b_n\}$ does not converge, then $\{a_n b_n\}$ does not converge.
- (ii) If $\{a_n\}$ and $\{b_n\}$ are monotonically increasing, then $\{a_n b_n\}$ is increasing.

2.94 ¶. Let $a_n > 0 \forall n$ and $a_{n+1}/a_n \rightarrow L$. Show that

- (i) $a_n \rightarrow 0$ if $0 \leq L < 1$,
- (ii) $a_n \rightarrow +\infty$ if $L > 1$.

2.95 ¶¶ Cesàro's theorems. Show that

- (i) $\liminf_n a_n \leq \liminf_n \frac{1}{n} \sum_{i=1}^n a_i \leq \limsup_n \frac{1}{n} \sum_{i=1}^n a_i \leq \limsup_n a_n$,
 - (ii) $\liminf_n (a_n - a_{n-1}) \leq \liminf_n \frac{a_n}{n} \leq \limsup_n \frac{a_n}{n} \leq \limsup_n (a_n - a_{n-1})$.
- In particular, in case $\{a_n - a_{n-1}\}$ has limit, then

$$\lim_{n \rightarrow \infty} \frac{a_n}{n} = \lim_{n \rightarrow \infty} (a_n - a_{n-1}).$$

- (iii) If $a_n \geq 0 \forall n$,

$$\liminf_{n \rightarrow \infty} a_n \leq \liminf_{n \rightarrow \infty} \sqrt[n]{\prod_{i=1}^n a_i} \leq \limsup_{n \rightarrow \infty} \sqrt[n]{\prod_{i=1}^n a_i} \leq \limsup_{n \rightarrow \infty} a_n;$$

in particular, if $\{a_n\}$ has limit, then

$$\lim_{n \rightarrow \infty} \sqrt[n]{\prod_{i=1}^n a_i} = \lim_{n \rightarrow \infty} a_n.$$

2.96 ¶. Discuss the convergence of the following sequences

$$\begin{aligned} \frac{1}{n+1} + \frac{1}{n+2} + \cdots + \frac{1}{2n}, & \quad \frac{\log n!}{n}, \\ \sqrt[n]{\frac{n^n}{n!}}, & \quad \sin\left(\pi\sqrt{4n^2 + \sqrt{n}}\right). \end{aligned}$$

2.97 ¶. Show that

$$\begin{aligned} \lim_{n \rightarrow \infty} \sin nx \text{ exists iff } x = k\pi, \quad k \in \mathbb{Z}, \\ \lim_{n \rightarrow \infty} \cos nx \text{ exists iff } x = 2k\pi, \quad k \in \mathbb{Z}. \end{aligned}$$

[Hint: Using the double-angle formula show that, if $\sin nx \rightarrow L$, then either $L = 0$ or $L = \pm \frac{\sqrt{3}}{2}$; use then the addition formula for $\sin(n+1)x$ to produce a contradiction if $x \neq k\pi$.]

2.98 ¶ Pythagorean algorithm. Set $a_1 = b_1 = 1$, and for $n \geq 1$

$$\begin{cases} a_{n+1} = a_n + b_n, \\ b_{n+1} = 2a_n + b_n. \end{cases}$$

Show that $b_n/a_n \rightarrow \sqrt{2}$. [Hint: Show first that

- $a_n, b_n \geq 2 \forall n \geq 2$,

- $\{a_n\}$ and $\{b_n\}$ are both strictly increasing,
- $a_n, b_n \rightarrow \infty$,
- $b_n^2 - 2a_n^2 = (-1)^n$.]

2.99 ¶ Heron's algorithm. Let α be a positive number. Define recursively the sequence $\{x_n\}$ by

$$\begin{cases} x_0 = \alpha > 0, \\ x_{n+1} = \frac{1}{2} \left(x_n + \frac{\alpha}{x_n} \right). \end{cases} \quad (2.32)$$

Show that $x_n \rightarrow \sqrt{\alpha}$. Show also that the absolute error $\delta_n := x_n - \sqrt{\alpha}$, verifies $\delta_{n+1} \leq \frac{1}{2\sqrt{\alpha}} \delta_n^2$, that is

$$\delta_{n+p} \leq 2\sqrt{\alpha} \left(\frac{\delta_p}{2\sqrt{\alpha}} \right)^{2^n}, \quad \forall n \geq 0 \text{ and } p \geq 1.$$

[Hint: Observe that

- $x_{n+1} - \sqrt{\alpha} = \frac{1}{2x_n} (x_n - \sqrt{\alpha})^2 \geq 0$,
- $\{x_n\}$ decreases.]

2.100 ¶¶. Let c and x_0 be positive real numbers. Show that the sequence defined inductively by

$$x_{n+1} = \frac{1}{p} \left((p-1)x_n + \frac{c}{x_n^{p-1}} \right)$$

converges to $\sqrt[p]{c}$.

2.101 ¶¶ Logarithm-arccosin algorithm. Consider the sequence

$$s_{n+1} := s_n \sqrt{\frac{2s_n}{s_n + s_{n-1}}}, \quad n = 0, 1, \dots$$

- (i) Fix $x > 0$. Set

$$s_{-1} := \frac{1}{4} \left(x^2 - \frac{1}{x^2} \right), \quad s_0 := \frac{1}{2} \left(x - 1/x \right)$$

and prove that $s_n \rightarrow \log x$ as $n \rightarrow \infty$.

- (ii) For $x \in [-1, 1]$, set

$$s_{-1} := x\sqrt{1-x^2}, \quad s_0 := x$$

and prove that $s_n \rightarrow \arcsin x$ as $n \rightarrow \infty$.

[Hint: (i) Show that $s_n = S(2^{-n})$ where $S(h) := \frac{1}{2h}(x^h - x^{-h})$, (ii) show that $s_n = R(2^{-n})$ where $R(h) := \frac{\sin(ha)}{h}$ and $a = \arcsin x$.]

2.102 ¶¶. Given $a, b > 0$ set

$$\begin{aligned} A(a, b) &:= \frac{a+b}{2}, & \text{the arithmetic mean,} \\ G(a, b) &:= \sqrt{ab}, & \text{the geometric mean,} \\ H(a, b) &:= \left(\frac{a^{-1} + b^{-1}}{2} \right)^{-1} = \frac{2ab}{a+b}, & \text{the harmonic mean.} \end{aligned}$$

- (i) Show that the sequences $\{x_n\}$ and $\{y_n\}$ defined recursively by

$$\begin{cases} x_0 = a, y_0 = b, \\ x_{n+1} = A(x_n, y_n), \\ y_{n+1} = G(x_n, y_n) \end{cases}$$

converge to the same limit, called the *arithmetic-geometric mean* of a and b .

- (ii) Show that the sequences $\{x_n\}$ and $\{y_n\}$ defined recursively by

$$\begin{cases} x_0 = a, & y_0 = b, \\ x_{n+1} = H(x_n, y_n), \\ y_{n+1} = A(x_n, y_n) \end{cases}$$

both converge to the geometric mean $G(a, b)$ of a and b .

2.103 ¶¶ Stirling's formula. Set $a_n := \log n! - \frac{1}{2} \log n$. Show that

- (i) $\int_{3/2}^n \log x \, dx < a_n < \int_1^n \log x \, dx$,
 (ii) $\delta_n := \left(1 - \int_1^n \log x \, dx\right) - a_n$ is decreasing and bounded below, thus has a limit $\delta \in \mathbb{R}$,
 (iii) deduce that the rough formula holds

$$n! \simeq e^\delta n^n e^{-n} \sqrt{n}.$$

[Hint: Compare with Example 2.67 and Section 7.4.]

2.104 ¶¶ Singular perturbation. Let $f : \mathbb{R} \rightarrow \mathbb{R}$. Show that

- (i) for any $y \in \mathbb{R}$ and $n \in \mathbb{N}$ the function

$$(f(x) - y)^2 + \frac{1}{n} x^2$$

has a minimum point x_n ,

- (ii) while $\{x_n\}$ may be unbounded, the sequence $\{x_n/\sqrt{n}\}$ is bounded,
 (iii) the sequence of minimum values

$$(f(x_n) - y)^2 + \frac{1}{n} x_n^2$$

has the limit as $n \rightarrow \infty$.

Leonardo Pisano (1170–1250), called Fibonacci			
Johannes de Sacrobosco (1195–1256)	Campanus of Novara (1220–1296)	Jean Buridan (1295–1358)	Nicole d' Oresme (1323–1382)
Albert of Saxony (1316–1390)	Karl Feuerbach (1800–1834)	Johann Regiomontanus (1436–1476)	
Luca Pacioli (1445–1517)			
Scipione del Ferro (1465–1526)	Nicolaus Copernicus (1473–1543)	Rudolf Stiffel (1487–1561)	
Colin MacLaurin (1698–1746)			
Niccolò Fontana (1500–1557), called Tartaglia	Girolamo Cardano (1501–1576)	Lodovico Ferrari (1522–1565)	Rafael Bombelli (1526–1573)
			Christopher Clavius (1537–1612)
François Viète (1540–1603)	Simon Stevin (1548–1620)	Bartholomeo Pitiscus (1561–1613)	
John Napier (1550–1617)		Henry Briggs (1561–1630)	
Galileo Galilei (1564–1642)			
Paul Guldin (1577–1643)	Marin Mersenne (1588–1648)	Girard Desargues (1591–1661)	Bonaventura Cavalieri (1598–1647)
René Descartes (1596–1650)		Pierre de Fermat (1601–1665)	
Gilles de Roberval (1602–1675)	Evangelista Torricelli (1608–1647)	John Wallis (1616–1703)	Blaise Pascal (1623–1662)
Christiaan Huygens (1629–1695)			
Vincenzo Viviani (1622–1703)	Pietro Mengoli (1626–1686)	Isaac Barrow (1630–1677)	Robert Hooke (1635–1703)
			James Gregory (1638–1675)
Sir Isaac Newton (1643–1727)		Gottfried von Leibniz (1646–1716)	

Figure 2.12. A chronological table from Fibonacci to Newton and Leibniz.

3. Integer Numbers: Congruences, Counting and Infinity

In this chapter we collect a few complements to the theory of integers. In Section 3.1, after discussing *Euclid's algorithm*, and the *fundamental theorem of arithmetic*, we deal with *Euler's function* and some of its applications to public key cryptography. In Section 3.2 we introduce a few basic elements of *combinatorics*, that is, the calculus of arrangements of a finite number of objects. Finally, in Section 3.3, we illustrate the notion of *cardinality* (or number of elements) of a (not necessarily finite) set introducing some of the concepts involved in *Cantor's theory of infinity*.

3.1 Congruences

3.1.1 Euclid's algorithm

Any subset of the integers, which is bounded above, has a maximum (see, for example, Proposition 1.23). An easy consequence of this is *division in the context of integers*.

3.1 Proposition. *Let $a, b \in \mathbb{Z}$, $b \neq 0$. Then the number a uniquely decomposes as*

$$a = qb + r \tag{3.1}$$

with $q, r \in \mathbb{N}$ and $0 \leq r < |b|$.

Proof. Suppose $b > 0$ and let q be the largest integer not greater than a/b , which exists by (vi) Proposition 1.23. From $p \leq a/b < p+1$ we get for $r := a - qb$ that $0 \leq r < b$.

If $b < 0$, choose p as the smallest integer not smaller than a/b : from $p \geq a/b > p-1$ we get for $r := a - qb$ that $0 < r \leq -b$.

It remains to prove that the decomposition is unique. If $a = q_1b + r_1 = q_2b + r_2$ with $0 < r_1, r_2 < |b|$, then $(q_2 - q_1)b = (r_1 - r_2)$: that is, $r_1 - r_2$ is an integral multiple of b . Since $|r_1 - r_2| < b$, we conclude that $r_1 - r_2 = 0$ and in turn that $q_1 = q_2$. \square

In (3.1) a is called the *dividend*, b the *divisor*, q the *quotient* and r the *remainder*. We write

$$q =: a // b, \quad r =: a \bmod b,$$

hence $a = (a // b)b + a \bmod b$, $\forall a, b \in \mathbb{Z}$, $b \neq 0$.

Let $a, b \in \mathbb{Z}$ and $b \neq 0$; we say that b *divides* a or that b *is a divisor of* a , if a is a multiple of b , that is, $a = bq$ for some $q \in \mathbb{Z}$. In this case we write $b|a$. Of course $b|a$ if and only if $a \bmod b = 0$, and, if a is a multiple of b and b is a multiple of c , then a is a multiple of c . Moreover, if a and b are multiples of c , then $ax + by$ is a multiple of c , too, $\forall x, y \in \mathbb{Z}$. In particular a and b are multiples of c if and only if a and $a \bmod b$ are multiples of c .

a. The greatest common divisor

Let $a, b \in \mathbb{Z}$ be nonzero integers. The set of common divisors of both a and b is nonempty and bounded. The largest of those numbers is called the *greatest common factor* or the *greatest common divisor* of a and b , and is denoted by

$$\text{g.c.d.}(a, b).$$

In other words, $r \in \mathbb{Z}$ is the greatest common divisor to a and b if and only if

- a and b are multiples of r ,
- if a and b are multiples of s , then $s \leq r$.

Trivially $\text{g.c.d.}(a, b) = \text{g.c.d.}(b, a) = \text{g.c.d.}(|a|, |b|)$. Finally, we say that $a, b \in \mathbb{Z}$ are *prime to one another* or *coprime* if $\text{g.c.d.}(a, b) = 1$. A number p is said to be *prime* if $p > 1$ and p has no positive divisor except 1 and p .

3.2 ¶. Show that

- (i) $\text{g.c.d.}(a, b) = b$ if and only if b divides a .
- (ii) Let p be prime. Then either $\text{g.c.d.}(a, p) = p$, that is a is a multiple of p , or $\text{g.c.d.}(a, p) = 1$, that is a and p are coprime.

We mentioned in Chapter 1 Euclid's algorithm as a method of finding a common submultiple to two magnitudes, discovering that in general it generates a process that never ends. When applied to two nonzero integers a and b , Euclid's algorithm stops after a finite number of steps and produces the greatest common divisor of a and b .

3.3 Euclid's algorithm. Let a, b be positive integers with $a > b$. The algorithm consists in dividing the larger of the two numbers by the smaller, then the smaller by the remainder of the first division, then the remainder of the first division by the remainder of the second, and so on. Formally, we set $r_0 := a$, $r_1 := b$, and

```

#!/usr/bin/env python

def euclid(a, b):
    r1, r2 = a, b
    while r2 != 0:
        r1, r2 = r2, r1 mod r2
    return r1

if __name__ == '__main__':
    print euclid(168,14)

```

Figure 3.1. Euclid's algorithm in Python.

$$\begin{aligned}
 r_2 &:= r_0 \bmod r_1, \\
 r_3 &:= r_1 \bmod r_2, \\
 r_4 &:= r_2 \bmod r_3, \\
 &\dots
 \end{aligned}
 \tag{3.2}$$

Since $r_0 > r_1 > r_2 > r_3 > \dots$ are nonnegative integers, the process will terminate with remainder zero after a finite number of steps,

$$\begin{aligned}
 r_n &:= r_{n-2} \bmod r_{n-1}, \\
 r_{n+1} &:= r_{n-1} \bmod r_n = 0.
 \end{aligned}
 \tag{3.3}$$

We have

3.4 Theorem (Euclid). $r_n = \text{g.c.d.}(a, b)$.

Proof. Observe that if $a = bq + c$ and $c \neq 0$, then r divides a and b if and only if r divides b and c . Consequently $\text{g.c.d.}(a, b) = \text{g.c.d.}(b, c)$. Iterating this observation along the algorithm, we get

$$\text{g.c.d.}(a, b) = \text{g.c.d.}(r_0, r_1) = \dots = \text{g.c.d.}(r_{n-2}, r_{n-1}) = r_n.$$

□

3.5 Remark. Euclid's algorithm can be written as

$$\begin{cases}
 r_0 = a, \ r_1 = b, \\
 q_k = r_{k-1} // r_k, \\
 r_{k-1} = q_k r_k + r_{k+1},
 \end{cases}$$

for $k = 1, 2, \dots, n$ as long as $r_k > 0$, i.e., until $r_{n+1} = 0$. Notice that $q_n \geq 2$.

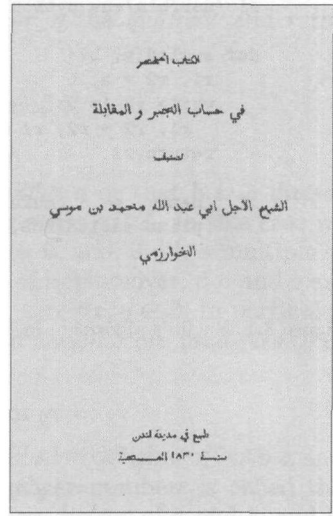
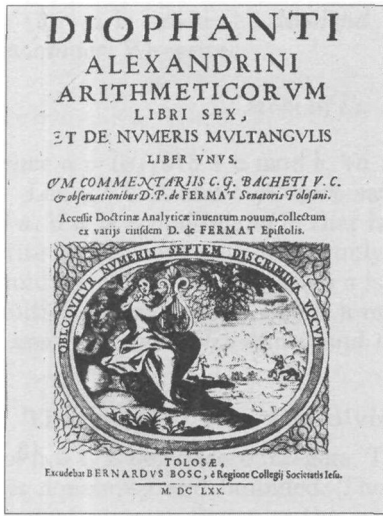


Figure 3.2. The frontispiece of the *Works* by Diophantus of Alexandria (200–284) with comments of Claude Gaspar Bochet de Méziriac (1581–1638) and observations by Pierre de Fermat (1601–1665), Tolosae 1670, and the frontispiece of *Algebra* by Abu al Khwarizmi (790–850).

3.6 Corollary. Let a, b and λ be integers with $a, b > 0$. Then

$$\text{g.c.d.}(\lambda a, \lambda b) = \lambda \text{g.c.d.}(a, b).$$

Proof. Let $\{r_n\}, \{s_n\}$ be the lists of remainders produced by Euclid's algorithm starting respectively with a, b and $\lambda a, \lambda b$. Since $\lambda a \bmod (\lambda b) = \lambda(a \bmod b)$, we deduce that $s_n = \lambda r_n \forall n$. Thus $r_{n+1} = 0$ if and only if $s_{n+1} = 0$, and

$$\text{g.c.d.}(\lambda a, \lambda b) = s_n = \lambda r_n = \lambda \text{g.c.d.}(a, b).$$

□

3.7 Corollary. We have

- (i) if c divides a and b , then $\text{g.c.d.}(a/c, b/c) = \text{g.c.d.}(a, b)/c$,
- (ii) $a/\text{g.c.d.}(a, b)$ and $b/\text{mcd}(a, b)$ are coprime,
- (iii) if $\text{g.c.d.}(a, b) = 1$ and a divides bc , then a divides c .

b. Integer solutions of first order equations

We discuss now the solvability in \mathbb{Z} of first order equations with integer coefficients, starting from the homogeneous case

$$ax + by = 0. \quad (3.4)$$

3.8 Proposition. *Let a, b be nonzero integers. Then the equation (3.4) is solvable in \mathbb{Z} and all solutions are given by the family of pairs (x_k, y_k) given by*

$$(x_k, y_k) := k \frac{1}{\max(a, b)}(b, -a), \quad k \in \mathbb{Z}. \quad (3.5)$$

Proof. Trivially all pairs in (3.5) solve the equation. Conversely suppose that (x, y) solves (3.4). Dividing by g.c.d. (a, b) we have

$$\frac{a}{\text{g.c.d.}(a, b)}x = -\frac{b}{\text{g.c.d.}(a, b)}y;$$

in particular $b/\text{g.c.d.}(a, b)$ divides the product on the left-hand side and, consequently x , since $a/\text{g.c.d.}(a, b)$ and $b/\text{g.c.d.}(a, b)$ are coprime (see, for example, Corollary 3.7). For some $k \in \mathbb{Z}$ we then have $x = k \frac{b}{\text{g.c.d.}(a, b)}$, and substituting into (3.4), we conclude. \square

Consider now the nonhomogeneous linear equation

$$ax + by = c, \quad c \neq 0.$$

3.9 Theorem (Bezout's theorem). *Let a, b, c be nonzero integers. The nonhomogeneous equation $ax + by = c$ is solvable if and only if g.c.d. (a, b) divides c .*

Proof. Suppose $x, y \in \mathbb{Z}$ solve $ax + by = c$. Trivially g.c.d. (a, b) has to divide c . Conversely, suppose that g.c.d. (a, b) divides c . We shall just prove that there exist integers x and y such that $ax + by = \text{g.c.d.}(a, b)$.

Consider the following recurrence scheme, known as the *generalized Euclid's algorithm*,

$$\left\{ \begin{array}{l} r_0 = a, \quad r_1 = b, \\ r_{k+1} = -q_k r_k + r_{k-1}, \quad q_k := r_{k-1} / r_k, \\ x_0 = 1, \quad x_1 = 0, \\ x_{k+1} = -q_k x_k + x_{k-1}, \\ y_0 = 0, \quad y_1 = 1, \\ y_{k+1} = -q_k y_k + y_{k-1}, \end{array} \right.$$

for $k := 1, 2, \dots, n$ as long as $r_n > 0$. Of course the r_k 's are the remainders of Euclid's algorithm, and it is easily seen by induction that $ax_k + by_k = r_k$ $\forall k = 0, 1, \dots, n$. Then Euclid's theorem yields

$$\text{g.c.d.}(a, b) = r_n = ax_n + by_n$$

as required. Going back to the equation $ax + by = c$, a solution is then given by

```
#!/usr/bin/env python

def euclid2(a, b):
    # assume a, b>0, a>b.
    r1, r2 = a, b
    x1, x2 = 1, 0
    y1, y2 = 0, 1
    while r2 <> 0:
        q, r = divmod(r1, r2)
        x1, x2 = x2, x1 - q*x2
        y1, y2 = y2, y1 - q*y2
        r1, r2 = r2, r
    return r1, x1, y1

if __name__ == '__main__':
    a, b = 1224, 12*17
    c, x, y = euclid2(a, b)
    print 'gcd=' , c
    print x, a, y, b
    print a*x+b*y
```

Figure 3.3. A Python implementation of the generalized Euclid's algorithm that computes a solution of $ax + by = \text{g.c.d.}(a, b)$.

$$(x, y) = \frac{c}{\text{g.c.d.}(a, b)}(x_n, y_n).$$

□

3.10 Remark. We also remark that Euclid's algorithm is quite efficient. Suppose that we perform two successive divisions

$$\begin{aligned} r_2 &= r_0 \bmod r_1, \\ r_3 &= r_1 \bmod r_2, \end{aligned} \tag{3.6}$$

then $r_3 < r_1/2$. In fact, either $r_2 \leq r_1/2$ and $r_3 < r_2 \leq r_1/2$, or $r_2 > r_1/2$, that is $2r_2 > r_1$, hence $q_3 = 1$ and again $r_3 = r_1 - r_2 < r_1/2$.

Euclid's algorithm in (3.2) requires at most $n + 2$ divisions in order to stop at a zero remainder, thus, if we denote by p the integral part of $(n + 1)/2$, we have in the worst case

$$1 \leq r_{n-1} \leq r_{n-2}/2 \leq \cdots \leq b/2^p,$$

i.e., $p \leq \log_2 b$. Therefore $(n + 1)/2 \leq p + 1/2 \leq \log_2 b + 1/2$, that is, $n \leq 2 \log_2 b$. We then conclude that *Euclid's algorithm stops after at most 2 times the number of digits in the binary representation of b* . For an optimal estimate, see Corollary 8.9.

3.11 ¶. Show that $\text{g.c.d.}(a, b)$ is the minimum of the set

$$A := \left\{ n \in \mathbb{N} \mid n > 0 \text{ and } n = ax + by, \text{ for some } x, y \in \mathbb{Z} \right\}.$$

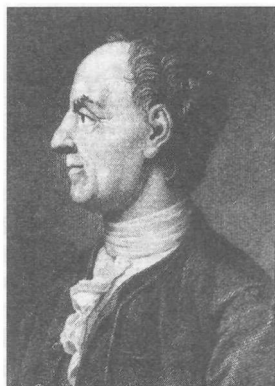


Figure 3.4. Pierre de Fermat (1601–1665) and Leonhard Euler (1707–1783).

3.1.2 Prime factorization

3.12 Theorem (The fundamental theorem of arithmetic). *Every integer $n \geq 2$ decomposes as a product of primes, and, apart from rearrangement of factors, such a decomposition is unique.*

Proof. We proceed by induction on n . If n is prime there is nothing to prove; otherwise, if p is the smallest of its divisors, p has to be prime as $n = mp$. Since $m < n$, by induction m and in turn n are a product of primes.

To prove uniqueness of the factorization, recall that by Corollary 3.7 if p is prime and p divides nm , then p divides n or m . Suppose $p_1 p_2 \cdots p_r = q_1 q_2 \cdots q_s$ where $p_1, \dots, p_r, q_1, \dots, q_s$ are prime. For each $j = 1, \dots, r$, p_j and the factors q_i are coprime, thus necessarily p_j is one of the q_i . It follows that $r \leq s$, and, changing the p 's with the q 's, we get $r = s$, and, apart from rearrangement, $p_i = q_i$ for all $i = 1, \dots, r$. \square

If we arrange the factors of the decomposition of $n \in \mathbb{N}$ in increasing order, we obtain that *every integer decomposes uniquely as*

$$n = p_1^{\alpha_1} p_2^{\alpha_2} p_1^{\alpha_3} \cdots p_k^{\alpha_k}$$

where $2 \leq p_1 < p_2 < \cdots < p_k$ are prime, and $\alpha_1, \dots, \alpha_k \geq 1$.

3.13 Theorem (Second Euclid's theorem). *The number of primes is infinity.*

Proof. Suppose p prime. Let $2, 3, 5, \dots, p$ be the set of primes up to p ; and let

$$n := 2 \cdot 3 \cdots (p-1)p + 1.$$

Then n is not divisible by any prime between 2 and p . On the other hand n has a decomposition in primes; therefore either n is prime or divisible by a prime between p and n . In either case there is a prime greater than p . \square

3.14 ¶. Show that g.c.d. (a, b) is the product of factors common to a and b , i.e., is the *greatest common factor* to a and b .

How do we identify the list of primes?

The simplest method consists in relying upon the definition of prime numbers. If $n = ab$, then a and b cannot exceed \sqrt{n} ; therefore any number that is not prime is divisible by a prime p which does not exceed \sqrt{n} . In order to decide whether p is prime, it suffices then to divide n by all primes less than \sqrt{n} . Notice that the method allows us to factorize p if p is not prime. In principle all is fine, but the method is impracticable even for numbers that are not very large: to conclude we need to do roughly \sqrt{n} divisions: if 10^{-9} seconds is the time needed to carry out a division, the time necessary to factorize a number of 100 digits is around $2^{50-20} = 2^{30}$ seconds, i.e., about 32 years.

A variant of the previous method is a procedure known as the *sieve of Eratosthenes* that allows us to find all primes not greater than n once we know all primes smaller than \sqrt{n} . It works as follows. Suppose we have a list $\{p_{s(n)}\}$ of all primes less than \sqrt{n} . From $p_{s(n)} + 1, p_{s(n)} + 2, \dots, n$ we strike out successively all multiples of $p_1, p_2, \dots, p_{s(n)}$ (up to n). The remaining numbers are all primes. In fact, if q is one of these numbers, it is not divisible by any of the primes less than \sqrt{n} , and consequently less than n . Also the sieve of Eratosthenes, though requiring multiplications instead of divisions, requires a number of multiplications of order \sqrt{n} . One says that *the computational complexity of the sieve of Eratosthenes is $O(2^{N/2})$* , $N := \log_2 n$ being the number of digits of n .

In the eighteenth century eventually the idea of looking for a complete description of primes was given up, and research moved toward a kind of statistical approach. The fundamental result in this direction is the remarkable *prime number theorem*, first conjectured by Adrien-Marie Legendre (1752–1833) and then proved by Jacques Hadamard (1865–1963) and Charles de la Vallée-Poussin (1866–1962)

3.15 Theorem (The prime number theorem). Let $\pi(x)$ denote the number of primes not greater than x . Then

$$\lim_{x \rightarrow +\infty} \frac{\pi(x)}{x / \log x} = 1.$$

There seems to be evidence that $x/\log x$ is a good approximation of $\pi(x)$: a celebrated conjecture, *Riemann's conjecture*, states that

$$\pi(x) = \frac{x}{\log x} + O\left(x^{\frac{1}{2}+\epsilon}\right)$$

for some $\epsilon > 0$, but it has not been proved or disproved up to now. Incidentally, observe the unexpected relation between prime numbers and the Euler's number stated by the prime number theorem.

3.1.3 Linear congruences

Let $p \in \mathbb{N}$. We say that a is congruent to b modulo p and write

$$a \equiv b \pmod{p}$$

if $a \bmod p = b \bmod p$. Equivalently $a \equiv b \pmod{p}$ if and only if $a - b = hp$ for some $h \in \mathbb{Z}$. It is easily seen that

- (i) if $a \equiv b \pmod{p}$ and $a' \equiv b' \pmod{p}$, then $a + a' \equiv b + b' \pmod{p}$ and $aa' \equiv bb' \pmod{p}$

and that congruence \pmod{p} is an equivalence relation, i.e., it is

- (i) REFLEXIVE. $a \equiv a \pmod{p}$,
- (ii) SYMMETRIC. if $a \equiv b \pmod{p}$, then $b \equiv a \pmod{p}$,
- (iii) TRANSITIVE. if $a \equiv b \pmod{p}$ and $b \equiv c \pmod{p}$, then $a \equiv c \pmod{p}$.

Equivalence classes of congruent numbers form a partition of all the integers into classes of numbers with the same remainder of the division by p . These classes can therefore be represented by the remainders of the division by p . Formally, one defines the set of remainder classes modulo p as

$$\mathbb{Z}_p := \{0, 1, 2, \dots, p-1\}$$

and the map $x \rightarrow x \bmod p$, whose image is \mathbb{Z}_p , yields a way to introduce the *sum* and the *product modulo p* in \mathbb{Z}_p by

$$a +_p b := (a + b) \bmod p \quad a \cdot_p b := (ab) \bmod p.$$

Congruences turn out to be quite important in everyday life. Here we confine ourselves to basic facts. From the results on the solvability in \mathbb{Z} of linear equations, we readily infer

3.16 Proposition. Consider the equation in $x \in \mathbb{Z}$,

$$ax \equiv c \pmod{p}. \tag{3.7}$$

- (i) If $c = 0$, then all solutions of (3.7) are given by the family of integers $\{x_k\}$ given by

$$x_k = \frac{p}{\text{g.c.d.}(a, p)} k, \quad k \in \mathbb{Z}.$$

- (ii) If $c \neq 0$, (3.7) has a solution if and only if $\text{g.c.d.}(a, p)$ divides c , and all the solutions are given by the sequence of integers $\{x_k\}$

$$x_k = \frac{p}{\text{g.c.d.}(a, p)} k + \frac{c}{\text{g.c.d.}(a, p)} \bar{x}, \quad k \in \mathbb{Z},$$

where \bar{x} is such that $a\bar{x} + p\bar{y} = 1$ for some $\bar{y} \in \mathbb{Z}$, and can be computed by the generalized Euclid's algorithm.

In doing computations with congruences modulo a composite number, the following theorem is quite useful.

3.17 Theorem (Chinese remainder theorem). *Let p_1, p_2, \dots, p_k be coprime.*

(i) $x \in \mathbb{Z}$ is a solution of the system

$$\begin{cases} x \equiv 0 \pmod{p_1}, \\ \dots \\ x \equiv 0 \pmod{p_k}, \end{cases}$$

if and only if $x \equiv 0 \pmod{p_1 p_2 \cdots p_k}$.

(ii) For any $(b_1, b_2, \dots, b_k) \in \mathbb{Z}^k$, the system

$$\begin{cases} x \equiv b_1 \pmod{p_1}, \\ x \equiv b_2 \pmod{p_2}, \\ \dots \\ x \equiv b_k \pmod{p_k}, \end{cases} \quad (3.8)$$

is solvable in $x \in \mathbb{Z}$.

More precisely, for any $i = 1, \dots, k$, denote by M_i the product of all primes $\{p_i\}$ but p_i , and let a_i be such that $M_i a_i \equiv 1 \pmod{p_i}$. Then $x := \sum_{i=1}^k M_i b_i a_i$ is a solution of (3.8), and two solutions of (3.8) differ by a multiple of $p_1 p_2 \cdots p_k$.

Proof. (i) It suffices to observe that, if p and q are coprime, then a is a multiple of both p and q if and only if a is a multiple of pq , and proceed inductively.

(ii) For any $i = 1, \dots, k$, observe that $M_j \equiv 0 \pmod{p_i}$ for $j \neq i$, and, since M_i and p_i are coprime, there exists $a_i \in \mathbb{Z}$ such that

$$M_i a_i \equiv 1 \pmod{p_i}.$$

Consequently

$$x \equiv \sum_{j=1}^k a_j b_j M_j \equiv a_i b_i M_i \equiv b_i \pmod{p_i}.$$

Finally, (i) implies that the difference of two solutions of (3.8) is a multiple of $p_1 p_2 \cdots p_k$. \square

Of particular relevance in the theory of congruences is the multiplicative subgroup \mathbb{Z}_p^* of the nonzero elements of \mathbb{Z}_p , and the *discrete exponential map* $x \rightarrow a^x \pmod{p}$ from \mathbb{Z}_p into \mathbb{Z}_p^* . As we can see in the table in Figure 3.5, we get $a^6 \equiv 1$ for all $a \neq 0$ in \mathbb{Z}_7 . This is a general fact first observed by Pierre de Fermat (1601–1665)

a	a^1	a^2	a^3	a^4	a^5	a^6
1	1	1	1	1	1	1
2	2	4	1	2	4	1
3	3	2	6	4	5	1
4	4	2	1	4	2	1
5	5	4	6	2	3	1
6	6	1	6	1	6	1

Figure 3.5. The table of $a \rightarrow a^n$ for \mathbb{Z}_7 .

3.18 Theorem (Fermat minor theorem). *If p is prime, then $a^{p-1} \equiv 1 \pmod{p} \forall a \neq 0$.*

The original proofs of many of the claims of Fermat were never found. The following proof is due to Euler (1739).

Proof. By the binomial theorem (see, for example, Example 2.6),

$$(1+a)^p = a^p + \sum_{k=1}^{p-1} \binom{p}{k} a^k + 1,$$

where $\binom{p}{k} := \frac{p(p-1)(p-2)\cdots(p-k+1)}{k!}$. Since p is prime and k is less than p , all binomial coefficients $\binom{p}{k}$, $k = 1, \dots, p-1$ are multiples of p . Therefore

$$(1+a)^p \equiv a^p + 1 \pmod{p} \quad \forall a \in \mathbb{Z}.$$

Using the previous equality, it is not difficult to show by induction on a that $a^p \equiv a \pmod{p}$. In fact, trivially $1^p \equiv 1 \pmod{p}$, and, if $b^p \equiv b \pmod{p}$, then

$$(b+1)^p \equiv b^p + 1 \equiv b + 1 \pmod{p}.$$

Finally, if $a \neq 0$, $\text{g.c.d.}(a, p) = 1$, consequently, choosing $x \in \mathbb{Z}$ such that $ax \equiv 1 \pmod{p}$, we conclude

$$a^{p-1} \equiv a^p \equiv ax \equiv 1 \pmod{p}.$$

□

A different proof was given then by James Ivory (1765–1842) in 1808 and presented again by Lejeune Dirichlet (1805–1859) in 1828.

A different proof of Theorem 3.18. If $a \neq 0$ is a multiple of p , the theorem is trivial. If a is not divisible by p , the numbers

$$a, 2a, 3a, \dots, (p-1)a$$

are not pairwise congruent modulo p . After rearrangement they are then congruent to $1, 2, \dots, p-1$. Thus

$$a \cdot 2a \cdot 3a \cdots (p-1)a \equiv 1 \cdot 2 \cdot 3 \cdots (p-1) \pmod{p},$$

that is,

$2^2 - 1 =$	$3 =$	3
$2^4 - 1 =$	$15 =$	$5 \cdot 3$
$2^6 - 1 =$	$63 =$	$7 \cdot 9$
$2^8 - 1 =$	$255 =$	$51 \cdot 5$
$2^{10} - 1 =$	$1023 =$	$11 \cdot 63$
$2^{12} - 1 =$	$4095 =$	$13 \cdot 315$

Figure 3.6. The values of $2^{n-1} - 1$ for $n = 3, 5, 7, 9, 11$ and 13 .

$$1 \cdot 2 \cdot 3 \cdots (p-1) a^{p-1} \equiv 1 \cdot 2 \cdot 3 \cdots (p-1) \pmod{p}.$$

Since $2 \cdot 3 \cdots (p-1)$ is not divisible by p , (v) yields

$$a^{p-1} \equiv 1 \pmod{p}$$

and in conclusion $a^p \equiv a \pmod{p}$. □

3.19 Pseudo-primes. One says that p is a pseudo-prime if $a^p \equiv a \pmod{p}$ for all $a \in \mathbb{Z}$. Of course primes are pseudo-primes, but there exist pseudo-primes that are not prime: they are called *Carmichael's numbers*, and the smallest is 561. Carmichael's numbers are quite rare, only 255 are not greater than 10^8 . Therefore it is likely that a number that is chosen randomly and verifies Fermat's test, is prime with probability close to 1. This means that the test $a^{p-1} \equiv 1 \pmod{p} \forall a \in \mathbb{Z}$, is a good indication that p is prime.

Fermat's test with $a = 2$ was actually the genesis of Fermat's theorem. Consider the table in Figure 3.6. It shows that $2^{n-1} - 1$ is divisible by n if n is 3, 5, 7, 11, 13. After many calculations Gottfried von Leibniz (1646–1716) conjectured that $2^{n-1} - 1$ is divisible by n if and only if n is prime. Fermat's theorem proves that 2^{n-1} is divisible by n if n is prime, but F. Edouard Lucas (1842–1891) in 1819 showed that the converse is not true. He showed that

$$\begin{cases} 2^{340} \equiv 1 \pmod{11}, \\ 2^{340} \equiv 1 \pmod{31}, \end{cases}$$

being that $2^5 \equiv -1 \pmod{11}$ and $2^5 \equiv 1 \pmod{31}$. Consequently $2^{340} \equiv 1 \pmod{341}$, i.e., $2^{340} - 1$ is divisible by $341 = 11 \cdot 31$, which is not prime. In recent years, the probability has been computed that a number n satisfies Fermat's test with $a = 2$ but is not prime. It turned out that this probability is quite small: for a randomly chosen number of, say, 200 bits, it is of the order $2.6 \cdot 10^{-8}$.

3.1.4 Euler's function ϕ

Given an integer $n \geq 2$, we denote by $\phi(n)$ the number of integers not greater than and prime to n (and therefore greater than 2), and we set

A^1	A^2	A^3	A^4	A^5	A^6	A^7	A^8	A^9	A^{10}	A^{11}	A^{12}	A^{13}	A^{14}
1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	4	8	1	2	4	8	1	2	4	8	1	2	4
3	9	12	6	3	9	12	6	3	9	12	6	3	9
4	1	4	1	4	1	4	1	4	1	4	1	4	1
5	10	5	10	5	10	5	10	5	10	5	10	5	10
6	6	6	6	6	6	6	6	6	6	6	6	6	6
7	4	13	1	7	4	13	1	7	4	13	1	7	4
8	4	2	1	8	4	2	1	8	4	2	1	8	4
9	6	9	6	9	6	9	6	9	6	9	6	9	6
10	10	10	10	10	10	10	10	10	10	10	10	10	10
11	1	11	1	11	1	11	1	11	1	11	1	11	1
12	9	3	6	12	9	3	6	12	9	3	6	12	9
13	4	7	1	13	4	7	1	13	4	7	1	13	4
14	1	14	1	14	1	14	1	14	1	14	1	14	1

Figure 3.7. The map $x \rightarrow A^x \bmod 15$.

$\phi(0) = \phi(1) = 1$. The function $\phi : \mathbb{N} \rightarrow \mathbb{N}$ defined this way is called *Euler's function* ϕ . Clearly,

$$\phi(p) = p - 1 \quad \text{if } p \text{ is prime.}$$

Suppose that p and q are coprime. Since p and q have no common divisors, we get that *Euler's function is multiplicative*, i.e.,

$$\phi(pq) = \phi(p)\phi(q) \quad \text{if } p, q \text{ are coprime.} \quad (3.9)$$

In particular

$$\phi(pq) = (p-1)(q-1) \quad \text{if } p, q \text{ are two distinct primes.}$$

Now we shall compute $\phi(p^k)$, p prime. Noticing that $z \geq 2$ divides p^k if and only if z is a divisor of p , we compute

$$\begin{aligned} \#\{x \geq 2 \mid \gcd(x, p^k) \neq 1\} &= \#\{h \mid 2 \leq hp < p^k\} \\ &= \#\{h \mid 1 < h < p^{k-1}\} = p^{k-1} - 1, \end{aligned}$$

and consequently

$$\phi(p^k) = p^k - 1 - (p^{k-1} - 1) = p^k \left(1 - \frac{1}{p}\right). \quad (3.10)$$

The unique factorization of any number in primes, (3.9) and (3.10) finally yield

3.20 Theorem (Euler). If $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$ is the decomposition in primes of n , $n \geq 2$, then

$$\phi(n) = n \left(1 - \frac{1}{p_1}\right) \left(1 - \frac{1}{p_2}\right) \cdots \left(1 - \frac{1}{p_k}\right).$$

We also prove

3.21 Theorem (Euler). If a and n are coprimes, then

$$a^{\phi(n)} \equiv 1 \pmod{n}.$$

Proof. Observe that, a and n being coprimes, x is coprime with n if and only if ax is coprime with n . Moreover ax and ay are congruent modulo n if and only if x and y are congruent modulo n . In other words, the map $x \rightarrow ax \pmod{n}$ is a bijection of the set

$$E := \left\{ x \in \{1, \dots, n-1\} \mid x \text{ and } n \text{ are coprime} \right\},$$

into itself. In particular

$$\prod_{x \in E} x = \prod_{x \in E} (ax) = a^{\phi(n)} \prod_{x \in E} x,$$

that is,

$$a^{\phi(n)} \equiv 1 \pmod{n}.$$

□

3.1.5 RSA Cryptography

Cryptography deals with the confidential transmission of data. Basically the sender *codes* the message M to a new object $C = f(M)$ by an injective map f defined on all possible messages, so that the addressee can *decode* the message by the inverse map, $M = f^{-1}(C)$. The function f is called the *cryptographic function*. For several reasons, for instance because any cryptographic function becomes less secure by use, it is important to change it from time to time and consider instead a *cryptographic system*, that is a family $\{f_k\}_{k \in K}$ of cryptographic functions indexed by a parameter called a *key*. Moreover, it is convenient to consider the cryptographic system as public and to ground the confidentiality of the system on the key.

One of the oldest ways of communicating secretly reduces to choosing openly a cryptographic system and then exchanging secretly a common key to code and decode messages. We then get a *bilateral communication*, that is the subjects are both senders and addressees. Among the recent algorithms for confidential connections between computers based on a common key, let us quote the AES (Advanced Encryption Standard) system, which is public, fast, and well discussed in the literature.

Exchanging a key is quite simple when the two subjects can meet in a safe place to physically exchange the key. Otherwise, they have to find a

safe channel of communication to send the key, but this leads us back to the starting point.

In 1976 Diffie and Hellman proposed a method that, in principle, enables people A (Alice) and B (Bob) to exchange safely a secret key, using an unsafe channel. It is based on modulo p arithmetic and on the different times necessary to compute the *discrete exponential*

$$x \rightarrow y := a^x \bmod p$$

and the *discrete logarithm*, that is, to solve the inverse problem of finding $x \in \mathbb{Z}$ such that

$$a^x = y \bmod p.$$

In fact, on the one hand the modular power operation can be computed with few multiplications. Writing x in binary form, that is $x = \sum_{i=0}^n x_i 2^i$, $x_i = 0$ or 1 , it is easy to see that the computation of $M^x \bmod n$ resolves in shifts and a number of multiplications of the order of the number of bits of x . On the other hand, the best known algorithm to solve $a^x \bmod p$ needs $2^{N^{1/3}}$ multiplications.¹ For large N , the second operation cannot be performed in a few years even using very powerful computers.

The procedure is as follows. Suppose that Alice and Bob want to exchange a numerical key k through a nonsafe public channel. First Alice and Bob openly choose a prime number p that is large enough and a number between 2 and $p - 1$. Then Alice chooses a number x that she keeps to herself and openly sends Bob $C := a^x \bmod p$. Analogously, Bob chooses a number y that he, too, keeps to himself and openly sends Alice $D := a^y \bmod p$. Alice and Bob are now able to build the common key by combining their secret number with the datum received from the other user: Alice computes

$$D^x \bmod p = a^{yx} \bmod p$$

and Bob computes

$$C^y \bmod p = a^{xy} \bmod p.$$

The two keys are identical. Nobody else will be able to recover x or y starting from p , a , C and D , without inverting the modular exponential.

In 1978 Rivest, Shamir, and Adelman proposed an algorithm that is now used in many programs for the so-called *public key cryptography*. Let us start with the following generalization of Fermat's theorem that is a key point in the RSA cryptographic system (see, for example, Figure 3.7).

3.22 Theorem. *Let p, q be two distinct primes and let $n := pq$. Then for all x and $k \in \mathbb{Z}$,*

$$a^{1+k\phi(n)} \equiv a \pmod{n},$$

$\phi(n)$ being the Euler's function, $\phi(n) := (p-1)(q-1)$.

¹ Well, if $a^x < p$, then the answer is immediate.

Proof. If a is not a multiple of p , by Fermat's theorem we have $a^{p-1} \equiv 1 \pmod{p}$, hence

$$a^{1+k\phi(n)} \equiv a \pmod{p}.$$

If a is a multiple of p , then $a^r \equiv 0 \equiv a \pmod{p}$ for all r , in particular $a^{k\phi(n)} \equiv 0 \equiv a \pmod{p}$, thus $a^{1+k\phi(n)} \equiv a \pmod{p}$. In all cases we infer

$$a^{1+k\phi(n)} \equiv a \pmod{p}.$$

Similarly

$$a^{1+k\phi(n)} \equiv a \pmod{q},$$

and, by the Chinese remainder theorem, $a^{1+k\phi(n)} \equiv a \pmod{n}$. \square

3.23 Corollary. Let p, q be two distinct primes and let $n := pq$. Let e, d be such that $ed \equiv 1 \pmod{\phi(n)}$. Then the maps

$$x \rightarrow x^d \pmod{n}, \quad y \rightarrow y^e \pmod{n}$$

from \mathbb{Z}_n into \mathbb{Z}_n are each the inverse of the other.

Proof. In fact $ed = 1 + k\phi(n)$, hence, by Theorem 3.22

$$(x^e \pmod{n})^d \equiv x^{ed} \equiv x^{1+k\phi(n)} \equiv x \pmod{n}.$$

\square

Assume that A (Alice) wants to communicate a secret to B (Bob). First Bob does the following:

- (i) he selects two primes p, q , and computes the *modulus* $n := pq$, and the Euler's function $\phi(n) := (p-1)(q-1)$,
- (ii) he selects a third prime e and computes d such that $ed \equiv 1 \pmod{\phi(n)}$, e.g., by the generalized Euclid algorithm,
- (iii) he publishes as *public key* the pair (n, e) and keeps as a *private key*, his secret, the pair (n, d) .

Assume that Alice (or anybody else) wants to send a confidential message to Bob. Firstly, she gets the public key (n, e) of Bob, then codes the messages as a number $M < n$, and finally she sends Bob $C := M^e \pmod{n}$. Bob is then able to recover M by computing $C^d \pmod{n}$, according to Corollary 3.23.

Besides the ability of Bob to decode the message, it is worth discussing the possibilities for an attacker to read the message, or, worse, to find the secret key (n, d) . We make only a few remarks.

- To decide if a given random number of around 256 digits is prime, Fermat's test with base $a = 2$ is usually sufficient, even if not totally secure. As a by-product, Fermat's test is fast, being based on modular powers. Recently, a fast algorithm of order $O(N^7)$ $N = \log_2 p$ was found that decides if a number p is prime or not. The algorithm does not exhibit a prime factor of p in case p is not prime, however.

- As we have already seen, computing $a^x \bmod n$ uses a number of multiplications of order $O(b)$, b being the number of bits of x . Assuming that the public and the secret exponents are between 0 and n , we conclude that the computational complexity of coding and of decoding the message is of order $O(N)$, $N := \log_2 n$.
- Computing $n := pq$ requires one multiplication, while finding the factors p, q from n , while in principle feasible, requires too much work for numbers $n = pq$ of, say, 256 bits. The sieve of Erathostenes is too slow requiring $2^{N/2}$ multiplications, $N = \log_2 n$, and even the better methods of factorization based on the LLL algorithm of Lenstra, Lenstra and Lovasz, predict a number of multiplications of order $2^{2.88N^{1/3}(\log N)^{2/3}}$, roughly 2^{55} for numbers n of 256 bits. Of course, this estimate holds for products of primes randomly chosen, while for special products the time of factorization can be shorter. Concluding, the choice of the primes p and q is somewhat critical, but generally speaking, choosing p, q in a random way with around 128 bits each, makes it practically infeasible to find in a short time the factors p, q from pq , and therefore impossible to get $\phi(n)$ and d from n this way in a short lapse of time.
- The RSA algorithm is clearly less secure than the factorization of integers, since it publishes also the number e . In fact a clever attack can be made to find d if d is small enough.

3.24 Theorem (Wiener, 1981). *Let p, q be primes with $p < q < 2p$. Let $n = pq$, and let $0 \leq e, d < \phi(n)$ be such that $ed \equiv 1 \pmod{\phi(n)}$. If*

$$d < \frac{1}{\sqrt{6}} \sqrt[3]{n},$$

then one can find d with an algorithm of computational complexity of $O(N)$, $N := \log_2 n$.

Proof. We have $0 \leq ed - 1 = k\phi(n) < d\phi(n)$, from which we infer $0 \leq k < d$ and $\text{g.c.d.}(k, d) = 1$. Moreover $p + q \leq 3\sqrt{n}$, therefore we can infer

$$\begin{aligned} \left| \frac{e}{n} - \frac{k}{d} \right| &= \left| \frac{de - kn}{dn} \right| \\ &= \frac{|de - k\phi(n) + k(\phi(n) - n)|}{dn} = \frac{|1 + k(\phi(n) - n)|}{dn} \\ &\leq \frac{k}{d} \frac{(n - \phi(n))}{n} < \frac{3}{\sqrt{n}} \end{aligned}$$

and

$$\left| \frac{e}{n} - \frac{k}{d} \right| < \frac{1}{2d^2},$$

since $1/\sqrt[3]{n} < 1/(\sqrt{6}d)$. Therefore k/d is one of the reduced continuous fractions of e/n , see Theorem 8.35. Since they can be all computed easily with $O(N)$ complexity by Euclid's generalized algorithm, the proof is concluded. \square

- A low exponent e , especially $e = 2, 3$, can be trouble, too, but the published flaws are far from a total break. Choices of e with around 17 bits are usually made.
- The RSA systems yields a way to make confidential communications from many to one. For a communication that involves two people, or two computers, the AES encryption standard, or other algorithms based on a common key, are preferred since they are by far faster than RSA. Thus RSA is often used for the only purpose of transmitting a secret key; then the actual communication is encrypted by the AES or similar algorithms. In this respect, the confidentiality of the transmitted message by RSA has to be analyzed, too. Despite 25 years of use of RSA, very few results are known on this subject. The basic question, whether inverting the RSA coding function is computationally equivalent to the factorization of $n := pq$, seems today a largely open issue.

The RSA algorithm can be useful also for authentication purposes. Assume that Bob codes a given public message with his secret key. Then anyone else can decode the coded version of the original message by Bob's public key, rediscovering the original message. In principle nobody can code the original message in such a way that the coded message, if decoded by Bob's public key, agrees to the original, unless the coding was done by Bob's secret key. This way, Bob can authenticate himself. However, the lack of mathematical evidence of the security of the RSA algorithm for authentication purposes and several documented breaks on some of the actual implementations, confines the RSA digital signature scheme to applications that need a mild form of authentication.

3.2 Combinatorics

We recall that a set X is said to be *finite* if there is a one-to-one correspondence of X with a subset of integers of the type $\{1, 2, \dots, n\}$. In this case, the *number n of elements* of X is called the *cardinality* of X and denoted by $|X|$ or $\#X$. In other words, every finite set X has $|X|$ elements and it can be ordered by given indices $1, 2, \dots, |X|$ to its elements.

Starting from one or several finite sets we can construct new sets either by *selecting* or *arranging* some of their elements or by taking *unions*, *intersections* or *products*. One refers to the procedures that allow computation of the number of elements of these new sets as *combinatorics*. This is a fascinating branch of mathematics with many applications, for instance, in engineering and social sciences. Here we confine ourselves to a few basic concepts.

3.2.1 Samples, mappings and subsets

Given n objects, how many different ways of selecting k objects from n do we have? In other words we want to compute the cardinality of the set of the selections of k objects out of n . The question is of course vague unless we specify

- whether the objects are all of the same type or how many objects belong to each type,
- the procedure according to which we make the choice,
- how we count the selected configurations.

a. Ordered samples and mappings

3.25 Definition. A list, or an ordered sample, $\{x_j\}_{j=0,1,\dots,k-1}$ of size k from a set X , an ordered k -sample in short, is an ordered selection of k elements from X .

Two lists $\{x_j\}$ and $\{y_j\}$ are equal if $x_j = y_j \forall j = 0, 1, \dots, k-1$, that is, if they contain the same elements arranged in the same order.

An ordered k -sample can be obtained by selecting its elements one after the other in many ways: for instance, we make the selection of each object from the entire population, so that the same element can be drawn more than once (*sampling with replacement*), or, an element once chosen is removed from the population (*sampling without replacement*). In the first case we are using *arrangements with repetitions*, in the second case *arrangements without repetitions*.

3.26 Arrangements without repetitions. What is the number of ordered k -samples without replacement from a population of n objects, called also *arrangements without repetitions* or *k-permutations* of n distinct objects? We have n choices for the first element, $n-1$ for the second, \dots , $(n-k+1)$ for the k -th. Consequently the number of ordered k -samples without replacement from n objects is

$$D_n^k := |\mathcal{D}_{n,k}| = n(n-1)(n-2)\cdots(n-k+1), \quad 1 \leq k \leq n.$$

For convenience we define also $D_n^0 := D_0^0 := 1$.

3.27 Permutations. An ordered n -sample from n objects is called a *permutation*. The set of permutations \mathcal{P}_n of a set of n elements has the cardinality

$$P_n := |\mathcal{P}_n| = D_n^n = n(n-1)\cdots 3 \cdot 2 \cdot 1 = n!.$$

We agree that the empty set has only one possible permutation, so that $P_0 := 1 = 0!$.

3.28 Arrangements with repetitions. In the case of ordered samples with replacement, the number of choices for the first element of a list as well as for the others is always n , thus the cardinality of the set of ordered k -samples with replacement from n objects, called also *permutations with repetitions of k objects from n* , is

$$D_n^{*k} := |\mathcal{D}_{n,k}^*| = |X^k| = |X|^k = n^k, \quad 1 \leq k \leq n.$$

We also set for convenience $D_n^{*0} := D_0^{*0} := 1$.

3.29 Maps. Ordered k -samples from n -elements with repetitions can be of course identified with the maps $f : \{1, 2, \dots, k\} \rightarrow \{1, \dots, n\}$, since every such map is defined by the list of its values

$$(f(1), f(2), \dots, f(k)).$$

From 3.28 the next proposition follows.

Proposition. Denote by $\mathcal{F}(X, Y)$ the set of all maps from X to Y . If $\#X = k$ and $\#Y = n$, then $\mathcal{F}(X, Y)$ has n^k elements.

Example. Given $A \subset X$, the *characteristic function* of A is defined by

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

The correspondence $A \subset X$ with the map $\varphi_A : X \rightarrow \{0, 1\}$ is clearly one-to-one. Therefore subsets of X are as many as the maps from X into $\{0, 1\}$, that is $\mathcal{P}(X) = 2^n$.

3.30 Injective maps. Also k -samples without repetitions from n objects can be of course identified with maps, the *injective* maps from $\{0, 1, \dots, k-1\}$ to $\{0, 1, \dots, n-1\}$. Consequently from 3.26 and 3.27 we get

Proposition. Denote by $\mathcal{I}(X, Y)$ the set of maps from X to Y that are injective. If $\#X = k$ and $\#Y = n$, then

$$\#\mathcal{I}(X, Y) = D_n^k = n(n-1)(n-2) \cdots (n-k+1)$$

In particular the number of bijective maps from X into itself is

$$\#\mathcal{I}(X, X) = k!$$

b. Nonordered samples and subsets

We often sample k elements from n objects, but the actual order of the elements in the resulting arrangement is unimportant, that is, two samples may be considered equal if they contain the same elements, irrespectively of the order. We then speak of *nonordered samples*.

3.31 Nonordered samples without replacement. As we have seen, the number of ordered k samples with replacement from n objects is D_n^k (see, for example, 3.26). Since we have $k!$ different ordered samples of the same k objects, the number of nonordered k -samples without replacement is

$$C_n^k := \frac{n(n-1) \cdots (n-k+1)}{k!} = \binom{n}{k}, \quad 1 \leq k \leq n.$$

We also set for convenience $C_n^0 = C_0^0 = 1$. Nonordered k -samples without replacement from n objects are also called *k-combinations* of n distinct objects.

The binomial coefficients are defined for all $\alpha \in \mathbb{R}$ and $k \in \mathbb{N}$ as

$$\binom{\alpha}{k} := \frac{\alpha(\alpha-1)(\alpha-2) \cdots (\alpha-k+1)}{k!}$$

and the following formulas hold

$$\begin{aligned} \binom{n}{k} &= \frac{n!}{k!(n-k)!} & \forall k, n \in \mathbb{N}, \\ \binom{-\alpha}{k} &= (-1)^k \binom{\alpha+k-1}{k} & \forall k \in \mathbb{N}, \alpha \in \mathbb{R}, \\ \binom{n}{0} &= \binom{n}{n} = 1, \\ \binom{n}{1} &= \binom{n}{n-1} = n, \text{ and } \binom{n}{k} = \binom{n}{n-k} & \forall k, 1 \leq k \leq n, \\ \binom{n}{k} &= \frac{n}{k} \binom{n-1}{k-1}, & \forall k = 1, 2, \dots, n, \\ \binom{\alpha}{k} &= \binom{\alpha-1}{k-1} + \binom{\alpha-1}{k}. \end{aligned}$$

The last formula is called *Pascal's formula*.

3.32 Subsets of finite sets. A nonordered k -sample without replacement from a set X of n elements, is merely a subset $A \subset X$ of k elements. Therefore, from 3.31 we get

Proposition. Let $\#X = n$. The number of subsets A of X with k elements is

$$\left| \left\{ A \in \mathcal{P}(X) \mid |A| = k \right\} \right| = C_n^k = \binom{n}{k}.$$

In particular, the number of subsets of X , including the empty set, i.e., the cardinality of $\mathcal{P}(X)$, is

$$1 + \sum_{k=1}^n \binom{n}{k} = \sum_{k=0}^n \binom{n}{k} = (1+1)^n = 2^n,$$

on account of the binomial theorem.

c. Ordered lists

3.33 Definition. Let X be an ordered set by \leq . A monotonic k -list $\{x_i\} \subset X$ is also called an ordered list.

3.34 Increasing lists. Let X be an ordered finite set. It is not restrictive to assume that $X = \{1, 2, \dots, n\}$. Let us compute the number L_n^k of the increasing lists of k numbers between 1 and n , i.e., of the type

$$\{h_1, h_2, h_3, \dots, h_k\}, \quad h_i < h_{i+1} \quad \forall i = 1, \dots, k-1.$$

Thinking of k lists from n objects as maps from $\{1, \dots, n\}$ to $\{1, \dots, n\}$, it is easy to see that ordered k -lists correspond to the strictly increasing maps. Since there is a unique way of listing k numbers in an increasing order, ordered k -lists are equal in number to the subsets of k elements of X , that is

$$L_n^k = C_n^k = \binom{n}{k}.$$

3.35 Nondecreasing lists. Let us compute the number L_n^{*k} of nondecreasing ordered lists of k objects from n , i.e., of k -uples $\{h_i\}$ with $h_i \leq h_{i+1} \quad \forall i = 1, \dots, k-1$. This time the elements of a list $\{h_i\}$ are not necessarily distinct. However, we can associate in a one-to-one fashion to each such nondecreasing k -list from n objects a strictly increasing k -list from $n+k-1$ objects by means of the map ϕ defined as

$$\Phi(h_1, h_2, \dots, h_k) := (h_1, h_2 + 1, h_3 + 2, \dots, h_k + (k-1)).$$

It is easily seen that ϕ maps the nondecreasing k -lists from $\{1, \dots, n\}$ to the increasing k -lists from $\{1, \dots, n+k-1\}$ and that Φ is one-to-one. Thus

$$L_n^{*k} = L_{n+k-1}^k = C_{n+k-1}^k = \binom{n+k-1}{k}.$$

Thinking of k lists from n objects as maps from $\{1, \dots, n\}$ to $\{1, \dots, n\}$, it is easy to see that nondecreasing k -lists correspond to nondecreasing maps from $\{1, \dots, k\}$ to $\{1, \dots, n\}$.

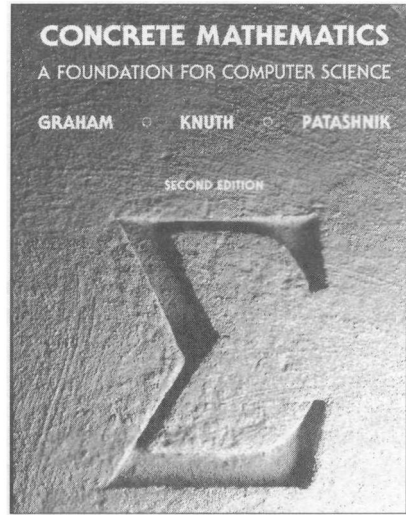
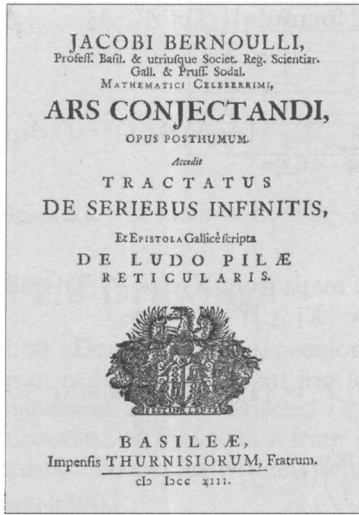


Figure 3.8. The frontispieces of the *Ars conjectandi* by Jacob Bernoulli (1654–1705) and a modern treatise about *discrete mathematics*.

3.36 Nonordered samples with replacement. Since the nonordered k -samples with replacement from n objects are clearly as many as the nondecreasing k -lists from the same objects, we conclude that the number of nonordered k -samples with repetitions, also called *k -combinations with repetitions* from n is

$$C_n^{*k} := \binom{n+k-1}{k} = (-1)^k \binom{-n}{k}, \quad 1 \leq k \leq n. \quad (3.11)$$

d. The formula of inclusion and exclusion

Let $A, B \subset \Omega$ be finite and disjoint subsets of Ω ; then we have $|A \cup B| = |A| + |B|$, and, more generally, in the case $A \cap B \neq \emptyset$,

$$|A \cup B| = |A| + |B| - |A \cap B|. \quad (3.12)$$

The formula (3.12) extends to the case of more than two subsets and is very useful in computing for example the probability of incompatible events. The reader will check that in the case of three subsets we have

$$|A_1 \cup A_2 \cup A_3| = |A_1| + |A_2| + |A_3| - |A_1 \cap A_2| - |A_2 \cap A_3| - |A_1 \cap A_3| + |A_1 \cap A_2 \cap A_3|.$$

Let A_1, A_2, \dots, A_n be finite subsets of Ω and $1 \leq k \leq n$. Since the intersection of k of the sets A_1, A_2, \dots, A_n is commutative, we index the intersection $A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}$ by an increasing list of indices $i_1 < i_2 < \dots < i_k$. It is not difficult to show that we have

3.37 Proposition (Inclusion-exclusion formula). *Let A_1, A_2, \dots, A_n be finite subsets of Ω . Then*

$$|A_1 \cup A_2 \cup \dots \cup A_n| = \sum_{k=1}^n (-1)^{k+1} \sum_{1 \leq i_1 < i_2 < \dots < i_k \leq n} |A_{i_1} \cap A_{i_2} \cap \dots \cap A_{i_k}|.$$

e. Surjective maps

Denote by $\mathcal{S}(X, Y)$ the family of *surjective* maps from X into Y . Trivially $|\mathcal{S}(X, Y)| = 0$ if $|X| < |Y|$ while in the case $|X| \geq |Y|$ we have

3.38 Proposition. *Let $X = \{1, 2, \dots, k\}$, $Y = \{1, 2, \dots, n\}$ and $n \leq k$. Then*

$$|\mathcal{S}(X, Y)| = \sum_{j=0}^n (-1)^j \binom{n}{j} (n-j)^k.$$

Proof. Let $\mathcal{F}(X, Y)$ be the family of all maps $f : X \rightarrow Y$ and, for $j = 1, \dots, n$, A_j denote the family of maps $f : X \rightarrow Y$ the ranges of which do not contain j . Trivially

$$\mathcal{F}(X, Y) = \mathcal{S}(X, Y) \bigcup \bigcup_j A_j,$$

hence the inclusion-exclusion formula yields

$$\begin{aligned} |\mathcal{S}(X, Y)| &= |\mathcal{F}(X, Y)| - |A_1 \cup \dots \cup A_j| \\ &= n^k - \sum_{j=1}^n (-1)^j \sum_{1 \leq i_1 < \dots < i_j \leq n} |A_{i_1} \cap \dots \cap A_{i_j}|. \end{aligned} \quad (3.13)$$

For every j -ple (i_1, \dots, i_j) with distinct elements, $A_{i_1} \cap \dots \cap A_{i_j}$ is the family of maps whose images contain at most $n-j$ elements; consequently

$$|A_{i_1} \cap \dots \cap A_{i_j}| = (n-j)^k$$

and

$$\begin{aligned} \sum_{1 \leq i_1 < \dots < i_j \leq n} |A_{i_1} \cap \dots \cap A_{i_j}| &= (n-j)^k \sum_{1 \leq i_1 < \dots < i_j \leq n} 1 \\ &= (n-j)^k \binom{n}{j}. \end{aligned} \quad (3.14)$$

The result then follows from (3.13) and (3.14). \square

k	0	1	2	3	4	5	6
D_5^{*k}	1	5	25	125	625	3125	0
D_5^k	1	5	20	60	120	120	0
C_n^k	1	5	10	10	5	1	0
C_n^{*k}	1	5	15	35	70	126	0

Figure 3.9. The values of D_5^{*k} , D_5^k , C_n^{*k} , C_n^k .

3.2.2 Drawings

3.39 Drawing in succession. The drawings in succession of k elements from n with replacement are as many as the ordered k -samples with replacement from n objects, $D_n^{*k} = n^k$ (see 3.28), while the drawings in succession of k elements from n without replacement are as many as the ordered k -samples without replacement from n objects, i.e., $D_n^k = \frac{n!}{(n-k)!}$ (see 3.26).

3.40 Simultaneous drawings. Making a simultaneous drawing of k objects from n is clearly the same as choosing a subset of k elements from n . Therefore the simultaneous drawings of k elements from n are as many as the subsets with k elements in a set with n elements, that is, $C_n^k = \binom{n}{k}$ (see, for example, 3.32).

3.41 Simultaneous drawings with repetitions. Suppose instead we have a population of infinitely many elements of type 1, infinitely many elements of type 2, ..., infinitely many elements of type n . The simultaneous drawings of k elements from this population are as many as the nonordered k -samples with replacement, that is, $C_n^{*k} = \binom{n+k-1}{k}$. The same result holds provided the initial population has at least k elements of each type.

The table in Figure 3.9 shows how results can be different.

3.2.3 Location problems

How many ways do we have of placing k balls into n cells? Again the question is quite vague unless we specify how to distinguish the resulting arrangements and the rules to fill the cells. In this respect we look at

- whether the balls are distinct,
- how many balls can be placed into a cell.

These kinds of problems arise typically in statistical mechanics. Several situations left vague by the previous description are quite relevant. The next examples describe some of them: they refer to the case of distinct cells.

3.42 Maxwell-Boltzmann statistics. The balls have a label which renders them distinct, moreover there is no limit to the number of balls that can be placed in one of the n cells. In this case the number of ways of placing k balls into n cells equals the number of maps $f: \{1, 2, \dots, k\} \rightarrow \{1, 2, \dots, n\}$, that is n^k .

If we instead require that only one of the distinct balls can be placed in a cell, we have as many possibilities as the injective maps $f: \{1, 2, \dots, k\} \rightarrow \{1, 2, \dots, n\}$ are, that is, $D_n^k = \frac{n!}{(n-k)!}$, $1 \leq k \leq n$.

3.43 Bose-Einstein statistics. The k balls are all black, and there is no limit on the number of balls we can place in a cell. In this case each arrangement can be regarded as a sequence of black balls separated by white balls, which represent the cells. We therefore have as many possibilities as the number of the subsets of $n-1$ elements in a set of $n-1+k$ elements, i.e.,

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

Another way of thinking is that any such arrangement be regarded as a nondecreasing map from $\{1, \dots, k\}$ into $\{1, \dots, n\}$.

3.44 Fermi-Dirac statistics. The k balls are nondistinct, and we cannot place more than one ball in each of the n -cells. In this case $1 \leq k \leq n$ and the number of possibilities are as many as the injective maps from $\{1, \dots, k\}$ into $\{1, \dots, n\}$, i.e., the subsets of k elements in $\{1, 2, \dots, n\}$, $\binom{n}{k}$.

3.45 ¶¶. A physical system consists of some identical particles. The total energy of the system is $4E_0$, $E_0 = \text{const} > 0$. Each particle may possess a level of energy kE_0 , $k = 0, 1, 2, 3, 4$, and a particle of energy kE_0 may occupy one of the $k^2 + 1$ states corresponding to this energetic level. How many different configurations, according to the energetic state of the particles, can the system assume? Answer the same question assuming that (a) at the energy level kE_0 there are $2(k^2 + 1)$ energetic states, (b) two particles are not allowed to occupy the same energetic state.

3.46 Example. How many lists of k integers exist with sum n ? In other words, what is the cardinality of the set

$$\{(x_1, x_2, \dots, x_k) \mid x_1 + \dots + x_k = n\}?$$

Interpret x_1, \dots, x_k as the number of nondistinct balls placed respectively in the cells $\{1, \dots, k\}$. Then the initial problem reads as: in how many ways can n nondistinct balls be placed in k cells? The answer is

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

The table in Figure 3.16 at the end of this chapter summarizes the different models of counting we have discussed in this section.

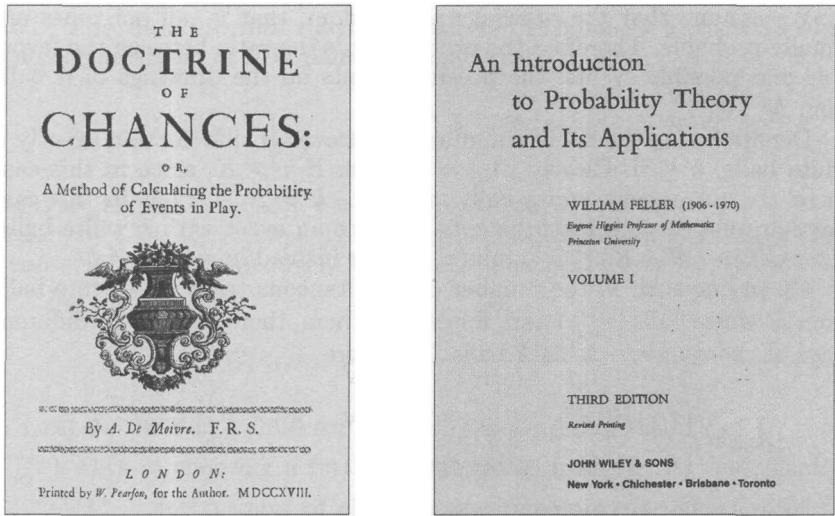


Figure 3.10. The frontispieces of the *Doctrine of Chances* by Abraham de Moivre (1667–1754) and a modern treatise on probability.

3.2.4 The hypergeometric and multinomial distributions

The birth of probability is dated back to Blaise Pascal (1623–1662) and to his correspondence with Pierre de Fermat (1601–1665) about a number of questions connected to the games of cards posed by the knight de Méré, who was a dogged gambler with mathematical velleity. The first published treatise, *De Ratiociniis in ludo aleae*, appeared in 1647 and is due to Christiaan Huygens (1629–1695); it was followed by the *Ars conjectandi* of 1713 by Jacob Bernoulli (1654–1705) and by *The Doctrine of Chances* of 1718 by Abraham de Moivre (1667–1754).

There are several definitions of *probability*. The first, and for this reason it is referred to as *classical*, is due to Blaise Pascal (1623–1662). The probability is the *ratio between the favorable events and all possible events, provided all events are equiprobable*. It is convenient to imagine the events as subsets A of a set Ω of the possible cases, and it is usual to assign probability 1 to the *certain event* Ω . The classical probability of the event A is then

$$P(A) = \frac{|A|}{|\Omega|}, \quad \forall A \subset \Omega,$$

this way reducing to a problem of counting.

3.47 The hypergeometric distribution. In this context a typical problem is the following. We are given a set X of N distinct balls: K of them are white and $N - K$ black. We simultaneously draw n balls from X . What is the probability for exactly k of them to be white?

We assume that the drawings are random, that is, all outcomes are equally probable. Therefore the probability is the ratio between the favorable and possible events, the possible events on the drawings of n balls from N , i.e., $\binom{N}{n}$.

Denote by A_k the set of simultaneous drawings that contain exactly k white balls, $k \leq n$. Clearly $|A_k| = 0$ either if $k > K$, since in this case there are not enough white balls, or if $n - k > N - K$, as in this case there are not enough black balls to produce an event with k white balls. If $\max(0, n - N + K) \leq k \leq \min(n, K)$, we instead have $|A_k| \neq \emptyset$.

Given one such k , the number of simultaneous trials of k white balls from K white balls is $\binom{K}{k}$ and, for each of them, there are $\binom{N-K}{n-k}$ different ways of choosing $n - k$ black balls. Therefore

$$|A_k| = \begin{cases} \binom{K}{k} \binom{N-K}{n-k} & \text{if } \max(0, n - N + K) \leq k \leq \min(n, K), \\ 0 & \text{otherwise,} \end{cases}$$

thus the probability of drawing n balls with exactly k white balls in the trial from our set X is

$$B(N, K, n)(k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}. \quad (3.15)$$

The (3.15) is called the *hypergeometric distribution*.

Notice that the sets A_k are pairwise disjoint and their union yields all possible drawings. Therefore we have

3.48 Proposition (Vandermonde formula). *Let $N, K, n \in \mathbb{N}$; then*

$$\sum_{k=\max(0, n-N+K)}^{\min(n, K)} \binom{K}{k} \binom{N-K}{n-k} = \binom{N}{n}.$$

Proof. In fact

$$\binom{N}{n} = |\Omega| = \sum_k |A_k| = \sum_{k=\max(0, n-N+K)}^{\min(n, K)} \binom{K}{k} \binom{N-K}{n-k}.$$

□

3.49 ¶. Write a proof of Vandermonde's formula using the identity $(1+a)^N = (1+a)^K (1+a)^{N-K}$.

3.50 ¶ Quality inspection. In an industrial production of 1000 items, 2% of them are defective. Choosing at random 25 items, what is the probability of finding two or more defective items?

3.51 The multinomial distribution. Let Ω be a set; a *partition* of Ω is a decomposition of Ω in pairwise disjoint sets A_1, A_2, \dots, A_p ,

$$\Omega = \bigcup_{i=1}^p A_i, \quad A_i \cap A_j = \emptyset \text{ for } i \neq j.$$

Let $n := |\Omega|$ and $k_i := |A_i|$, $i = 1, \dots, p$. Of course $k_1 + k_2 + \dots + k_p = n$. Denote by $C(k_1, k_2, \dots, k_p)$ the number of possible decompositions of Ω in p subsets drawing respectively k_1, k_2, \dots, k_p elements with $k_1 + k_2 + \dots + k_p = n$. We then have

$$C(k_1, k_2, \dots, k_p) = \frac{n!}{k_1! k_2! \dots k_p!}.$$

In fact, we have $\binom{n}{k_1}$ different ways of choosing k_1 elements of Ω , then $\binom{n-k_1}{k_2}$ ways of choosing a further k_2 elements from Ω , \dots , and, finally, $\binom{n-(k_1+k_2+\dots+k_{p-1})}{k_p}$ ways of choosing the remaining k_p elements of Ω . Therefore

$$\begin{aligned} C(k_1, k_2, \dots, k_p) &= \frac{n!}{k_1!(n-k_1)!} \frac{(n-k_1)!}{k_2!} \dots \frac{(n-(k_1+k_2+\dots+k_{p-1}))!}{k_p!} \\ &= \frac{n!}{k_1! k_2! \dots k_p!}. \end{aligned}$$

3.52 ¶. 52 cards are distributed to four players. How many hands are possible?

3.3 Infinity

3.3.1 The mathematical analysis of infinity

Already Galileo Galilei (1564–1642) remarked that the squares of natural numbers are *as many as* the natural numbers themselves. In *Discorsi intorno a due nuove scienze* he wrote

*Interrogando io ... quanti siano i numeri quadrati, si può con verità rispondere, loro esser tanti quanti sono le proprie radici, avvenga che ogni quadrato ha la sua radice, ogni radice il suo quadrato, né quadrato alcuno ha più d'una sola radice, né radice alcuna più d'un quadrato solo.*²

² Asked ... how many are the squared numbers, one can verily say they are as many as the square roots, in fact every square number has its square root and every square root its square, nor any square has more than a square root nor any square root has more than a square.

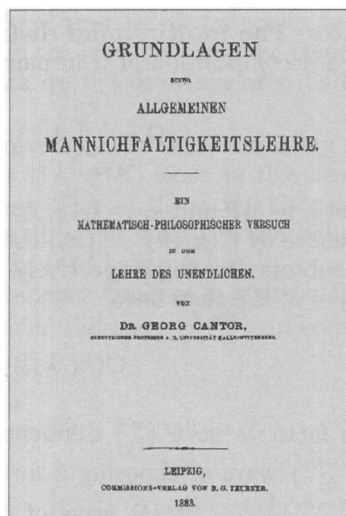
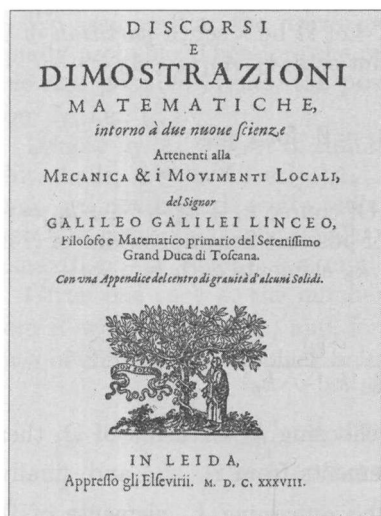


Figure 3.11. The frontispieces of the *Discorsi intorno a due nuove scienze* by Galileo Galilei (1564–1642) and a book by Georg Cantor (1845–1918) about infinity.

For a long time the only notion of infinity really accepted by mathematicians was Aristotle's notion of *potential infinity*, in the sense of never ending: the natural numbers as 0, 1 and so on is an example of potential infinity. But the use of *actual infinity* was either avoided or used as a source of contradictions, and according to the mathematical and philosophical Greek tradition: *infinitem actu non datur*.³

But the development of mathematics, especially in the eighteenth and nineteenth centuries, led mathematicians to confront themselves not only with the idea of potential infinity in the sense of infinite processes as in the infinitesimal calculus, but also with the need of understanding the structure of infinite sets.

a. Cardinality

At the end of the eighteenth century Georg Cantor (1845–1918), in a series of papers, set the foundations of the theory of sets and, in particular, analyzed the concept of infinity on the basis of the principle of one-to-one correspondence.

3.53 Definition. Two sets A and B are said to be equivalent or to have the same power or the same cardinality, and we write $\text{card } A = \text{card } B$ or $A \sim B$, if and only if they are in a one-to-one correspondence with each other.

It is easily seen that \sim is an equivalence relation, i.e., it is

³ Actual infinity is not given.

- (i) REFLEXIVE. $A \sim A$,
- (ii) SYMMETRIC. $A \sim B$ if and only if $B \sim A$,
- (iii) TRANSITIVE. If $A \sim B$ and $B \sim C$, then $A \sim C$.

The equivalence classes are called the *cardinals*: if A belongs to the equivalence class α , we say that α is the cardinality of A and we write $\alpha := |A|$ or $\alpha := \text{card } A$; and, existence of a cardinal α means the existence of a set A with $\text{card } A = \alpha$.

Sets A that have the same power of $\{0, 1, 2, 3, \dots, n-1\}$ are said to have cardinality n ; the empty set has cardinality 0 by definition. This way natural numbers become cardinals: they are called *finite cardinals*, all other are called *transfinite cardinals*.

If A and B are *disjoint* sets of cardinality α and β , the cardinality of $A \cup B$ and of $A \times B$ depends only on α and β and is denoted by $\alpha + \beta$ and $\alpha\beta$; in particular, if $\alpha = \alpha_1$ and $\beta = \beta_1$ then $\alpha + \beta = \alpha_1 + \beta_1$. The cardinality $\alpha + \beta$ agrees with the ordinary sum of integers if α and β are finite cardinals.

If $A \neq \emptyset$, the set of mappings from A into B is denoted by B^A , and its cardinality by β^α . If α and β are finite, then β^α is the ordinary power with integers, see Section 3.2.

The set of naturals \mathbb{N} is infinite and its cardinality is denoted by \aleph_0 (aleph, \aleph , is the first letter of the Hebrew alphabet). A set that has cardinality \aleph_0 , that is in one-to-one correspondence to \mathbb{N} , is called *denumerable* or *countable*.

Of course a set has cardinality \aleph_0 if and only if it is possible to *enumerate* it.

3.54 ¶. Show that

- (i) $\text{card } \{2n \mid n \in \mathbb{N}\} = \text{card } \{2n+1 \mid n \in \mathbb{N}\} = \aleph_0$,
- (ii) $\text{card } \{n \in \mathbb{N} \mid n \geq \bar{n}\} = \aleph_0$,
- (iii) $\bar{n} + \aleph_0 = \text{card } (\{0, 1, \dots, \bar{n}-1\} \cup \mathbb{N}) = \aleph_0$,
- (iv) $\aleph_0 + \aleph_0 = \aleph_0$,
- (v) $n \aleph_0 = \aleph_0$ for all $n = 1, 2, 3, \dots$.

[Hint: Show a bijection between the sets involved. To prove (iii) notice that a bijection $\{0, 1, \dots, n-1\} \times \mathbb{N} \rightarrow \mathbb{N}$ is given by $(i, k) \rightarrow i + nk$.]

3.55 Definition. Let α, β be cardinals. We say that

- (i) $\alpha \leq \beta$ if we can find sets A and B such that $A \subset B$, $\text{card } A = \alpha$ and $\text{card } B = \beta$.
- (ii) $\alpha < \beta$ if $\alpha \leq \beta$ and $\alpha \neq \beta$.

Trivially $\text{card } A \leq \text{card } B$ if $A \subset B$. If A is infinite and $B \subset A$ is finite, obviously, $A \setminus B$ is infinite. Consequently we can inductively choose $a_1 \in A$, $a_2 \in A \setminus \{a_1\}$, $a_3 \in A \setminus \{a_1, a_2\}$, and so on; therefore A contains a subset that has the same power of \mathbb{N} . We conclude

3.56 Proposition. A set A is infinite if and only if it contains a denumerable subset. A cardinal α is transfinite if and only if $\alpha \geq \aleph_0$.

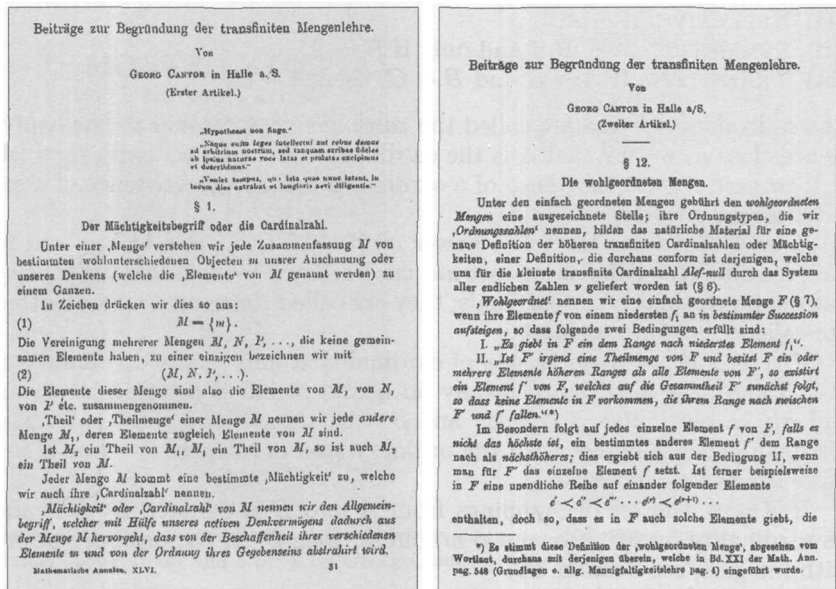


Figure 3.12. Two pages from two of Cantor's papers about infinity that appeared respectively in 1895 and 1897 in *Mathematische Annalen*.

If A is infinite, according to Proposition 3.56, we can write

$$A = A_1 \cup A_2 \quad \text{with} \quad \text{card } A_1 = \aleph_0 \text{ and } A_1 \cap A_2 = \emptyset.$$

Since A_1 has the same power as a strictly included subset, A has the same cardinality as its proper subset $B_1 \cup A_2$, hence we can state

3.57 Proposition. *A set is infinite if and only if it has the same power as one of its proper subsets.*

b. Cantor–Bernstein theorem

In principle two cardinals are not comparable. However the following theorem states that $\alpha = \beta$ if and only if $\alpha \leq \beta$ and $\beta \leq \alpha$.

3.58 Theorem (Cantor–Bernstein). *If A is equivalent to a subset of B and B is equivalent to a subset of A , then A and B are equivalent.*

Proof. Let $h : A \rightarrow B_1 \subset B$ and $k : B \rightarrow A_1 \subset A$ be the one-to-one correspondence between A and $B_1 := h(A) \subset B$ and between B and $A_1 := k(B) \subset A$, and let $A_2 := k(B_1)$. Writing $E \sim F$ for $\text{card } E = \text{card } F$, by assumption $A \sim B$, and, by construction, $B \sim A_2$; hence $A_2 \sim A$. The map $\varphi := k \circ h : A \rightarrow A_2$ is one-to-one; set $A_0 := A$ and

$$A_{n+2} := \varphi(A_n), \quad \forall n \geq 1.$$

Since $A_2 \subset A_1 \subset A_0$, we have $A_{n+1} \subset A_n \forall n \geq 0$, hence

$$A_{n+2} \sim A_n \quad \forall n$$

and

$$(A_n \setminus A_{n+1}) \sim \varphi(A_n \setminus A_{n+1}) = A_{n+2} \setminus A_{n+3}.$$

Since the sets $A_{2j} \setminus A_{2j+1}$, $j = 0, 1, \dots$, are pairwise disjoint, also the subsets

$$H := \bigcup_{j=0}^{\infty} (A_{2j} \setminus A_{2j+1}), \quad \text{and} \quad \varphi(H) = \bigcup_{j=1}^{\infty} (A_{2j} \setminus A_{2j+1})$$

have the same power. On the other hand trivially

$$A = H \cup (A_1 \setminus A_2) \cup \bigcap_{i=0}^{\infty} A_i =: H \cup L,$$

$$A_1 = \varphi(H) \cup (A_1 \setminus A_2) \cup \bigcap_{i=0}^{\infty} A_i =: \varphi(H) \cup L,$$

hence $A \sim A_1$, consequently $A \sim B$. \square

An immediate consequence of the Cantor–Bernstein theorem and of Proposition 3.56 is

3.59 Proposition. \aleph_0 is the first transfinite cardinal, i.e., $\alpha < \aleph_0$ if and only if α is finite.

We notice that Proposition 3.59 does not follow directly from Proposition 3.57 since a priori $\alpha < \aleph_0$ is not alternative to $\aleph_0 \leq \alpha$.

c. Denumerable sets

Since the subset of primes in \mathbb{N} is infinite, it is denumerable. Similarly the set

$$\{p^n \mid p \text{ prime}, n \in \mathbb{N}\}$$

is denumerable. Since A is in one-to-one correspondence with $\mathbb{N} \times \mathbb{N}$, we infer

$$\aleph_0 \cdot \aleph_0 = \aleph_0.$$

More generally, the set

$$\{p_1 p_2^2 \cdots p_n^n \mid p_1, p_2, \dots, p_n \text{ prime}\}$$

is denumerable, hence $\aleph_0^n = \aleph_0$. The previous relation can be inferred by means of the following procedure known as the *first Cantor diagonal method*. Let $I_n = \{a_i^n\}_{i \in \mathbb{N}}$ be a countable family of denumerable sets. Enumerating $\bigcup_n I_n$ as follows

$$a_1^1, a_1^2, a_1^3, a_2^1, a_2^2, a_2^3, a_3^1, a_3^2, a_3^3, a_4^1, \dots,$$

compare with Figure 3.13, we see that

$$\text{card } \bigcup_{n=1}^{\infty} I_n = \aleph_0.$$

3.60 ¶. Show that

- (i) \mathbb{Z}^n , $n \geq 1$, is denumerable;
- (ii) \mathbb{Q} is denumerable;
- (iii) the set of polynomials with integer coefficients is denumerable;
- (iv) a real number is said to be an *algebraic number* if it is a root of a polynomial with integer coefficients. Show that the set of algebraic numbers is countable.

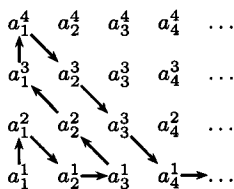


Figure 3.13. The first Cantor diagonal method.

d. The axiom of choice

Let A be a set; a *partial order* on A , denoted by \leq , is a relation on A with the properties

- (i) REFLEXIVE. $x \leq x \ \forall x \in A$,
- (ii) SYMMETRIC. if $x, y \in A$, $x \leq y$ and $y \leq x$, then $x = y$,
- (iii) TRANSITIVE. if $x, y, z \in A$, $x \leq y$ and $y \leq z$, then $x \leq z$.

Notice that we do not require that necessarily either $x \leq y$ or $y \leq x$. If this last property occurs, we say that \leq is a (*total*) *order* on A . On a tree, there is a natural partial order, in \mathbb{R} there is an order. The set of parts $\mathcal{P}(X)$ of a set X is partially ordered by the inclusion: If $A, B \subset X$, then $A, B \in \mathcal{P}(X)$, and we can say that $A \leq B$ in $\mathcal{P}(X)$ if and only if $A \subset B$.

Because of the Cantor–Bernstein theorem, the relation \leq defines a partial order on the family of cardinals. Does it define a total order? In other words, given two ordinals, is it true that either $\alpha \leq \beta$ or $\beta \leq \alpha$? The question is equivalent to the following: given a set A with $\text{card } A = \alpha$, and a cardinal β , such that $\beta \geq \alpha$ does not hold, can we construct a subset $B \subset A$ with $\text{card } B = \beta$? Of course the construction of B involves the choice of elements of A , and, in the case $\beta = \aleph_0$, we actually constructed a countable set B by induction, see Proposition 3.57.

In 1904, Ernst Zermelo (1871–1951) showed, but we are not going to present the proofs here, that the answer is positive, i.e., the following theorem holds.

3.61 Theorem. We have $\text{card } A \leq \text{card } B$ or $\text{card } B \leq \text{card } A$ for every pair of sets A and B if and only if we admit the following axiom of choice.

3.62 Axiom (Zermelo’s axiom). Let \mathcal{A} be a family of nonempty and pairwise disjoint sets. Then there exists a set C such that $C \cap A$ consists exactly of one element for each $A \in \mathcal{A}$.

Nowadays Zermelo’s axiom is widely accepted as one of the standard mathematical tools, though in the years many attempts have been made to avoid its use in many mathematical theories.

There are many equivalent ways of expressing Zermelo’s axiom and often the equivalence is not at all detectable at first sight.

3.63 Theorem. *The following claims are equivalent*

- (i) Zermelo's Axiom 3.62.
- (ii) AXIOM OF CHOICE. Let $\{X_i\}_{i \in I}$ be a family of nonempty sets with indices in a set I . Then there exists a function choice φ defined on each $i \in I$ such that $\varphi(i) \in X_i \forall i \in I$, that is, we can choose an element $x_i = \varphi(i)$ in each X_i and consider the set $\{x_i\}_{i \in I}$.
- (iii) CARTESIAN PRODUCT. The Cartesian product of the family $\{X_i\}_{i \in I}$, $\prod_{i \in I} X_i$, is empty if and only if one of the factors X_i is empty.

Suppose a partial order \leq is defined on A , and let $C \subset A$. In this situation we can easily introduce the notions of upper bound, supremum and maximum of C . An *upper bound* for C is an element m such that $c \leq m \forall c \in C$; the *supremum* of C is, if it exists, the lowest of the upper bounds m of C ; $c_0 \in A$ is the *maximum* of C if $c_0 \in C$ and $c \leq c_0 \forall c \in C$. Moreover, we say that c_0 is a *maximal element* for C if there is no $b \in A$ such that $a \leq b$ and $b \neq a$; finally, a totally ordered subset of C is called a *chain* of C .

With the previous definitions we can now formulate two other equivalent forms of Zermelo's axiom.

3.64 Theorem (well-ordering). *On every set X there is an order such that every nonempty subset has minimum.*

3.65 Theorem (Zorn's lemma). *Let A be a partial ordered set by \leq , and suppose that every chain of it has supremum. Then for every $a \in A$ there exists a maximal element $x \in A$ such that $a \leq x$.*

e. The power of the continuum

3.66 Theorem (Cantor). *Let A be a countable set. The family $\mathcal{P}(A)$ of subsets of A has the same power as the family 2^A of mappings $\varphi : A \rightarrow \{0, 1\}$, i.e., 2^{\aleph_0} , and it is strictly larger than \aleph_0 .*

Proof. The map that associates to each subset $E \subset A$ its characteristic function $\chi_E(x)$ (defined by $\chi_E(x) = 1$ if $x \in E$ and $\chi_E(x) = 0$ if $x \notin E$) clearly defines a bijection between $\mathcal{P}(A)$ and 2^A . It remains to show that $2^{\aleph_0} > \aleph_0$, that is, one cannot enumerate the family of sequences with values 0 and 1. We shall prove this using the *second Cantor diagonal method*. Suppose we are able to enumerate all sequences of 0 and 1. In this case we can form the table

$$\begin{array}{cccc} a_1^1 & a_2^1 & a_3^1 & \dots \\ a_1^2 & a_2^2 & a_3^2 & \dots \\ a_1^3 & a_2^3 & a_3^3 & \dots \\ \dots & \dots & \dots & \dots \end{array}$$

where the a_j^i are either 0 or 1. We now define a new sequence with values in $\{0, 1\}$ which is not listed in the previous table, a contradiction. For that, define

$$x_k = \begin{cases} 1 & \text{if } a_k^k = 0, \\ 0 & \text{if } a_k^k = 1. \end{cases}$$

Clearly $\{x_k\}$ does not agree with any of the lines in the table. \square

If we now observe that whenever $\text{card } A > \aleph_0$, $B \subset A$ with $\text{card } B = \aleph_0$, then $\text{card } A \setminus B = \text{card } A$, and if we represent the reals in a binary basis, we easily conclude that $\text{card } [0, 1] = 2^{\aleph_0}$. Also since $\tan(\pi(x - 1/2))$, $x \in]0, 1[$, is a bijection between $]0, 1[$ and \mathbb{R} , we infer that

$$\text{card } \mathbb{R} = 2^{\aleph_0}$$

i.e., 2^{\aleph_0} is the power of the continuum. Finally by the first Cantor diagonal method we have $\text{card } \mathbb{R}^n = 2^{\aleph_0}$, too. We can then summarize

3.67 Theorem. *The sets $[0, 1]$, $[0, 1]^n$, $n \geq 2$, and \mathbb{R}^n have all the power of the continuum.*

3.68 ¶. The real numbers that are not algebraic (see, for example, Exercise 3.60) are called *transcendental*. Show that the set of transcendental numbers has the power of the continuum.

The claim that the segment $[0, 1]$ and the n -dimensional cube have the same power deserves a few comments. The claim in Theorem 3.67 means that *there exists a one-to-one map between $[0, 1]$ and $[0, 1]^n$* . As paradoxical as it may appear, it says in particular that the concept of power or cardinality and of *dimension*, that is, in its intuitive form, the number of independent variables needed to describe a particular situation, are unrelated: for example it is not enough to describe an object in a one-to-one way with two parameters in order for it to be a surface. Actually the notion of dimension is related to more refined structures than just counting points, as, for instance, continuity.

f. The continuum hypothesis

More generally one shows that $2^\alpha > \alpha$ for any cardinal α . This way we can construct a hierarchy of transfinite cardinals

$$\text{card } \mathbb{N} < \text{card } \mathcal{P}(\mathbb{N}) < \text{card } (\mathcal{P}(\mathcal{P}(\mathbb{N}))) < \dots \quad (3.16)$$

The natural question of whether such a hierarchy exhausts all transfinite cardinals naturally arises.

The hypothesis that the cardinality of the continuum is the smallest nondenumerable cardinal, i.e., that no other cardinal lies between \aleph_0 and 2^{\aleph_0} is called the *continuum hypothesis*, while one refers to the assumption that the hierarchy in (3.16) exhausts all transfinite cardinals as the *generalized continuum hypothesis*.

In 1939 Kurt Gödel (1906–1978) showed that the generalized continuum hypothesis (in particular the continuum hypothesis) is consistent with

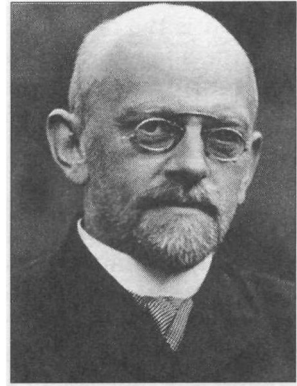
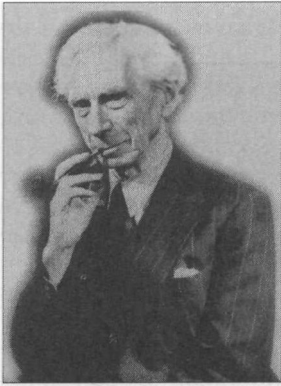


Figure 3.14. Bertrand Russell (1872–1970) and David Hilbert (1862–1943).

the standard axioms of set theory; in 1963 Paul Cohen (1934–) showed that also its negation (or the negation of the continuum hypothesis) is consistent with the same axioms. In other words, the continuum hypothesis is *independent* from the axioms of set theory and we can develop a set theory in which it is valid and a set theory in which it is not valid.

3.3.2 Some information on the theory of sets

In the second half of the eighteenth century mathematicians realized that a reasonable theory of sets was necessary for the development of mathematics. It is commonly agreed that the creator of the theory of sets was Georg Cantor (1845–1918) who developed the theory in many papers and made use of it in several contexts, and especially in the study of *cardinal* and *ordinal numbers*.

At the same time as Cantor, Gottlob Frege (1848–1925) developed a formal theory of the *higher order calculus of predicates*. This theory may be regarded as a theory of sets based on two axioms:

- AXIOM OF EXTENSIONALITY. Two sets are equal if they contain the same members.
- AXIOM OF ABSTRACTION. Given a predicate $p(x)$, there exists the set of x that satisfy $p(x)$.

Frege's axioms in connection with sets that are too large lead to several *paradoxes*. Cantor himself observed that the set of all sets should have a maximum cardinality K contradicting the fact that $2^K > K$. A similar observation had already been made by Cesare Burali-Forti (1861–1931) in connection with the theory of ordinals. In 1902 Bertrand Russell (1872–1970) observed that the axiom of abstraction is contradictory; in fact, if R is the set of all sets that are not members of themselves, then R is a member of itself.

What is the reason for paradoxes and how can we avoid them? A first reason was found by J. Henri Poincaré (1854–1912) and Bertrand Russell (1872–1970), and consists in the use of so-called *impredicative* notions, that is, in the use of quantifiers acting on all members of the set in order to define a new element. A first attempt to make up for this was the *theory of types* developed by Bertrand Russell (1872–1970) and Alfred N. Whitehead (1861–1947) in *Principia Mathematica*. However, excluding impredicative notions has some consequences, for instance the definition of the supremum for subsets of \mathbb{R} is impredicative. Another reason for the occurrence of paradoxes was seen by L. E. Brouwer (1881–1966) and the intuitionists in the *principle of excluded middle* (*tertium non datur*): either p or not p . They say that such a principle holds in correspondence of finite sets, but not in situations in which we use quantifiers on infinite sets. For the intuitionists the fact that “ $p(x)$ holds $\forall x$ ” does not hold does not imply that “there is x such that $p(x)$ does not hold.” They in fact interpret (or better pretend that one should interpret) the existence of x as the procedure or the exhibit of an x . On this basis the intuitionists started a program of reformulation of mathematics, that later on turned out to be of extreme relevance for information science, but doing that they also came to unsatisfactory conclusions such as, for instance, that every real function that exists in their sense is continuous.

Nobody, or hardly anybody, is willing to give up Cantor’s results, as Hilbert put it “no one will expel us from the paradise which Cantor created for us,” and Bertrand Russell describes Cantor’s work as “probably the greatest of which the age can boast.” However, in order to compare two cardinals α and β (i.e., say whether $\beta = \alpha$, $\alpha \leq \beta$ or $\beta \leq \alpha$) Cantor had to assume that every set can be well-ordered, a counter-intuitive claim.

In 1904 Ernst Zermelo (1871–1951) showed that *every set can be well-ordered*. In 1908 he analyzed the assumptions from which the theorem follows and which do not allow inference of the old paradoxes, though it does not answer the question of whether the new axioms would give rise to new paradoxes.

Zermelo gives up Frege’s axiom of abstraction (which is contradictory) and replaces it with *rules* that produce admissible sets (by means of union, intersections and powers) and with a weaker form of the axiom of abstraction, the

- AXIOM OF SEGREGATION. Given a set X and a predicate $p(x)$, there exists the subset $\{x \in X \mid p(x)\}$.

In the previous axiom p may be impredicative, i.e., may contain quantifiers on all X .

With these axioms, Zermelo was able to produce finite sets such as

$$\{\emptyset, \{\emptyset\}, \{\{\emptyset\}\}, \{\emptyset, \{\emptyset\}\}\}$$

but cannot produce infinite sets. For this reason Zermelo assumes also

- AXIOM OF INFINITY. There exists a set that contains the empty set and contains $\{x\}$ if it contains x .

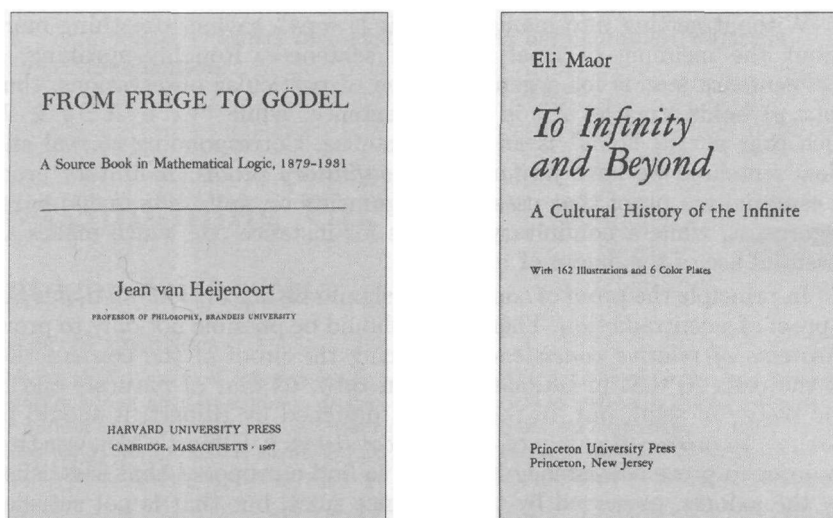


Figure 3.15. The frontispieces of a collection of sources of *Mathematical logic* and of a popular book about infinity.

Notice that the axiom of infinity states, in terms of sets, the existence of the natural numbers as an actual infinite set and not just as a potential infinite set. Finally, Zermelo states

- **AXIOM OF CHOICE.** *Given a family of disjoint nonempty sets $\{X_\alpha\}$, then there exists a set C which has as its members one and only one element from each X_α ,*

which is crucial for the proof of the well-ordering theorem.

The previous axioms, slightly modified by Abraham A. Fraenkel (1891–1965), are Zermelo–Fraenkel axioms of the theory of sets that allow one to prove the well-ordering theorem. Zermelo’s idea is: if we accept those claims, then we also have to accept the well-ordering theorem.

In contrast with the intuitionists, David Hilbert (1862–1943) started a *new program*. For Hilbert “forbidding a mathematician to make use of the principle of excluded middle is like forbidding an astronomer his telescope.” For Hilbert we need to distinguish between the *formalism*, which is *finitistic*, and *interpretations* of the formalism which may be *nonfinitistic*: for example, the calculus of polynomials is finitistic, but the interpretation of the formalism of polynomials as polynomial functions is *nonfinitistic*. Hilbert’s idea is then that formalism, being finitistic, always works, however, a part of it has a meaning which is accepted by everybody, i.e., *real sentences*, but *ideal sentences* may have a meaning which is not unanimously accepted, but, in any case, ideal sentences can be used to infer *real sentences*. From this point of view the fundamental question is that of the *consistency of the system* and the central question becomes the question of the *consistency of the Zermelo–Fraenkel axioms*.

Without getting into many details it is worth saying something more about the meaning of ideal and real sentences. Roughly speaking, a *real sentence* stands for a generalization of particular observations, thus " $p(x,y)$ holds $\forall x,y \in X$ " is a real sentence, while " $\forall x \in X \exists y \in X$ such that $p(x,y)$ holds" is an ideal sentence. Corresponding to real and ideal sentences we have *finitary* and *nonfinitary* proofs. A finitary proof is essentially a proof that uses only arguments on finite sets or inductive arguments, while a nonfinitary proof is for instance one which makes an essential use of the axiom of infinity.

In principle the proof of consistency should be finitary: for all d , d is not a proof of a contradiction. Therefore it should be possible not only to prove *theorems of relative consistency* (reducing the proof of the construction of the reals to that of rationals and, in turn, to that of naturals and of the theory of sets), but, in the context designed by Hilbert, it should be possible to prove a theorem of *absolute consistency*. Hilbert's idea was that in order to prove consistency it suffices to find a property, that is satisfied by the axioms, preserved by the inference rules, but that is not satisfied by a contradictory sentence. This way the proof of consistency could be carried out by induction.

Hilbert (as many other mathematicians) was worried by the fact that paradoxes, confined for the time to areas away from the kernel of mathematics, would enter the field of mathematical analysis, just refounded on nonfinitary arguments of set theory. On the other hand he firmly believed that every proof, even nonfinitary proofs, could be formally analyzed as a finite sequence of symbols on formulas, worked out according to precise syntactical rules (instead of as a flow of ideas, meanings and concepts), and consequently could be handled in a finitary way.

Hilbert's program reached a crisis when in 1931 Kurt Gödel (1906–1978) proved his celebrated *incompleteness theorems*, a consequence of which being that one can exhibit real sentences (in the sense of Hilbert) that can be proved only with nonfinitary means using the same axioms except the one asserting the existence of an infinite set. Later it was proved that one can exhibit a polynomial (in more than one variable) with integer coefficients without integer roots, a finitary claim, that however requires in an essential way the use of the axiom of infinity.

But there is more. Modulo coding numbers, Gödel proves that *the claim of consistency of the Zermelo–Fraenkel axioms is not provable* not only with finitary but even with nonfinitary means. However, the sum of knowledge acquired leads and transforms into the study of *formal systems*, that is into a new branch of mathematics, though, as Hermann Weyl (1885–1955) states,

the question of the ultimate foundations and the ultimate meaning of mathematics remains open; we do not know in what direction it will find its final solution or even whether a final objective answer can be expressed at all. "Mathematizing" may well be a creative activity of man, like language or music, of primary

originality,”⁴ whose historical decisions defy complete objective rationalization.

3.4 Summing Up

Integer arithmetic

Integral numbers

The integral division of integers leads naturally to the notions of *prime* and *coprime numbers* and of *greatest common divisor* of two numbers as well as to *Euclid’s algorithm* for finding the greatest common divisor of two numbers. An extension, *Euclid’s generalized algorithm*, produces a solution $(x, y) \in \mathbb{Z}^2$ of the equation $ax + by = \text{g.c.d.}(a, b)$, which allows computation of all solutions of the linear equation with integral coefficients

$$ax + by = c,$$

see *Bezout’s theorem*, Theorem 3.9.

- FUNDAMENTAL THEOREM OF ARITHMETIC. Every integer $n \geq 2$ decomposes as a product of primes and, apart from rearrangement of factors, that decomposition is unique.

Congruences

Bezout’s theorem solves *linear first order congruences modulo p* :

- $ax \equiv c \pmod{p}$ is solvable if and only if c is a multiple of $\text{g.c.d.}(a, p)$. In this case one is able to find all the solutions, see Proposition 3.16.
- $ax \equiv 1 \pmod{p}$ is always solvable with a unique solution $x \in \{0, \dots, p-1\}$ if p is prime. Thus the *ring of the remainders modulo p* , \mathbb{Z}_p , is a field if p is prime.
- CHINESE REMAINDER THEOREM. Given p_1, p_2, \dots, p_n coprimes, the system

$$\begin{cases} x \equiv b_1 \pmod{p_1}, \\ x \equiv b_2 \pmod{p_2}, \\ \dots \\ x \equiv b_n \pmod{p_n} \end{cases}$$

is solvable for any b_1, b_2, \dots, b_n , and two solutions differ by a multiple of $p_1 p_2 \dots p_n$. The Chinese remainder theorem is often used to solve $ax \equiv b \pmod{n}$ when n is a product of distinct primes.

A useful tool to analyze the multiplicative structure of the ring \mathbb{Z}_n , is the *exponential modular function* from \mathbb{Z}_n into \mathbb{Z}_n given by $x \rightarrow a^x \pmod{n}$. We have

- FERMAT’S MINOR THEOREM. If p is prime, then $a^p \equiv 1 \pmod{p} \forall a \in \mathbb{Z}_n, a \neq 0$.
- EULER’S THEOREM. Denote by $\phi(n)$ the number of integers $\leq n$ that are coprime with n . Then $a^{\phi(n)} \equiv 1 \pmod{n}$ for all a coprime with n .
- Let p and q be prime. Set $n := pq$ and let e and d be such that $ed \equiv 1 \pmod{\phi(n)}$. Then the two modular power maps from \mathbb{Z}_n into \mathbb{Z}_n given by

$$x \rightarrow a^e \pmod{n} \quad \text{and} \quad x \rightarrow a^d \pmod{n}$$

are one the inverse of the other.

The latter sentence is the foundation of the RSA public key cryptography.

⁴ And usefulness, we add.

Model	Arrangements	Drawings	Mappings	Locations	Statistical Physics
n	population	number of balls	cardinality of the range	cells	states
k	samples	drawn balls	cardinality of the domain	balls	particles
n^k	ordered k -samples with replacement from n	number of drawings in succession of k balls with replacement from n	number of mappings $\{1, \dots, k\} \rightarrow \{1, \dots, n\}$	number of ways of locating k distinct balls in n cells	Maxwell-Boltzmann statistics
$\frac{n!}{(n-k)!}$	ordered k -samples without replacement from n	number of drawings in succession of k balls without replacement from n		number of ways of locating k distinct balls in n cells, with at most one ball per cell	
$(-1)^k \binom{-n}{k}$	unordered k -samples with replacement from n			number of ways of locating k nondistinct balls in n cells	Bose-Einstein statistics
$\binom{n}{k}$	unordered k -samples without replacement from n	number of simultaneous drawings of k balls from n	number of injective maps from $\{1, \dots, k\}$ to $\{1, \dots, n\}$	number of ways of locating k nondistinct balls in n cells, with at most one ball per cell	Fermi-Dirac statistics

Figure 3.16. Samplings in different models of counting.

Combinatorics

The table in Figure 3.16 summarizes the numbers of ordered and nonorderd samples in the various models of counting: arrangements, sets and maps, drawings and locations.

Cardinals

Cardinality is a way to count elements in a set. Two sets have the same cardinality if there is a bijection between them, and cardinals are simply the equivalence classes of sets which are in a one-to-one correspondence.

One distinguishes sets with *finite cardinality*, or simply finite, that is the sets which are in a one-to-one correspondence with bounded sets of \mathbb{N} . The cardinality of such sets is just the number of elements they have. The other sets are called *infinite*, and their cardinals are said to be *transfinite*. Among these sets, the sets which are in a one-to one correspondence with \mathbb{N} are called *denumerable* or *countable*, and their cardinality is denoted by \aleph_0 . These sets are obviously infinite.

The inclusion relation between sets defines a relation on cardinals: $\alpha \leq \beta$ if and only if there exist sets A and B such that $\text{card } A = \alpha$, $\text{card } B = \beta$, and $A \subset B$. We also set $\alpha < \beta$ iff $\alpha \leq \beta$ and $\alpha \neq \beta$. The relation \leq between cardinals is obviously reflexive and transitive, while the symmetry is given by the

- CANTOR-BERNSTEIN THEOREM. Let α and β be two cardinals. If $\alpha \leq \beta$ and $\beta \leq \alpha$, then $\alpha = \beta$.

It then follows

- The relation \leq between cardinals is a partial order.
- Let $A \subset \mathbb{N}$. Then, either A is finite or it is in a one-to-one correspondence with \mathbb{N} . Equivalently, \aleph_0 is the first transfinite cardinal.
- \mathbb{Z} , \mathbb{Q} , and for $n \geq 1$, \mathbb{N}^n , \mathbb{Z}^n , \mathbb{Q}^n are denumerable.

At the beginning of the nineteenth century, Zermelo showed that the partial order relation \leq between cardinals actually is a total order, that is, given sets A and B we have either $\text{card } A \leq \text{card } B$ or $\text{card } B \leq \text{card } A$, provided we assume the following:

- ZERMELO'S AXIOM OF CHOICE. Let \mathcal{A} be a family of nonempty and pairwise disjoint sets. Then there exists a set C such that $C \cap A$ consists exactly of one element for each $A \in \mathcal{A}$.

Nowadays the axiom of choice is tacitly accepted, hence the possibility to compare different cardinals.

- CANTOR. Let $\alpha := \text{card}(A)$. Denote by 2^α the cardinality of the set of all maps $\varphi: A \rightarrow \{0, 1\}$. Then $2^\alpha > \alpha$. In particular $2^{\aleph_0} > \aleph_0$.
- $[0, 1] \subset \mathbb{R}$, \mathbb{R} and more generally, for every $n \geq 2$, \mathbb{R}^n have cardinality 2^{\aleph_0} . It is therefore impossible to distinguish sizes and “dimensions” by counting points.

3.5 Exercises

3.69 ¶. Find a number that is divisible by 7 and that, divided by 2, 3, 4, 5 or 6, yields always a remainder 1.

3.70 ¶. The *least common multiple* of two positive integers a and b is the least positive number that is divisible by both a and b . It is denoted by $\text{l.c.m.}(a, b)$. Show that

$$\text{l.c.m.}(a, b) \text{ g.c.d.}(a, b) = ab.$$

3.71 ¶. If p and q divide a and $\text{g.c.d.}(p, q) = 1$, then pq divides a .

3.72 ¶. Find the $\text{g.c.d.}(a, b)$ and the $\text{l.c.m.}(a, b)$ for each of the following pairs of integers:

$$(15000, 32768), \quad (46035, 47430), \quad (17795, 43291), \quad (2295, 1989).$$

3.73 ¶. Solve $ax + by = \text{g.c.d.}(a, b)$, $x, y \in \mathbb{Z}$ for the pairs (a, b) that follow:

$$(1542, 2102), \quad (2287, 442), \quad (1485, 1547), \quad (38, 127).$$

3.74 ¶. Show that p is prime if and only if $\text{g.c.d.}(a, p) = 1$ for all a , $2 \leq a < p$.

3.75 ¶. Let $d \in \mathbb{N}$, $d \geq 2$. Show that each $n \in \mathbb{N}$ can be uniquely represented as

$$n = a_0 + a_1d + a_2d^2 + \cdots + a_kd^k = \sum_{j=0}^k a_jd^j$$

with $0 \leq a_i \leq d-1 \ \forall i$, known as the *representation of n in bases d* .

3.76 ¶. Let a and b be coprime. Show that

$$\frac{1}{ab} = \frac{x}{a} + \frac{y}{b}$$

with $x, y \in \mathbb{Z}$.

3.77 ¶. Show that every rational number $r = p/q$, $p, q \in \mathbb{Z}$, $q \neq 0$, has a unique representation of the form

$$r = \frac{x_1}{p^{\alpha_1}} + \frac{x_2}{p^{\alpha_2}} + \cdots + \frac{x_k}{p^{\alpha_k}}$$

where $\alpha_1, \alpha_2, \dots, \alpha_k$ are integer coefficients, and p_1, p_2, \dots, p_k are distinct primes.

3.78 ¶. If p is prime, show that $(a+b)^p \equiv a^p + b^p \pmod{p}$.

3.79 ¶. Show that

- (i) n is divisible by 3 if and only if the sum of its digits (in base 10) is divisible by 3,
- (ii) n is divisible by 9 if and only if the sum of its digits (in base 10) is divisible by 9.
- (iii) $n = \sum_{j=0}^k a_j 10^j$ is divisible by 11 if and only if the alternate sum of its digits $a_0 - a_1 + a_2 - a_3 + \cdots + (-1)^k a_k$ is divisible by 11.

3.80 ¶¶. Show that for every $N > 1$ there exists N consecutive numbers none of which is prime. [Hint: If p is prime and $p > N$, consider the numbers $p!+2, p!+3, \dots, p!+p$.]

3.81 ¶¶. Deduce from the prime number theorem that, if p_n is the n -th prime number, then

$$\lim_{n \rightarrow \infty} \frac{p_n}{n/\log n} = 1.$$

3.82 ¶. Let $\{\alpha_k\}$ be a sequence of real numbers in binary representation. Cantor's diagonal procedure then produces a real number $\alpha \notin \{\alpha_k\}$. In particular, every sequence of algebraic numbers produces a nonalgebraic number.

3.83 ¶. Find the probability that two persons chosen at random were born on a Monday.

3.84 ¶. In how many different ways

- (i) can 8 persons be seated in 5 seats?
- (ii) can 5 persons be seated in 8 seats?

3.85 ¶ Poker. Find the probability for a poker hand to be three of a kind or a full house.

3.86 ¶ Méré paradox. Show that it is more probable to get at least one ace with four dice than at least one double ace in 24 throws of two dice.

3.87 ¶. Drawing successively with replacement c balls from n labeled from 1 to n , what is the probability of drawing k different balls?

3.88 ¶. From a box containing n distinct balls, what is the probability that a sample of size k , obtained with replacement, contains two equal balls?

3.89 ¶ Birthdays. What is the probability that in a population of n people the birthday of at least two people will fall on the same day, assuming equal probability for each day? Compute the probability for $n = 10, 25, 50$. Suppose $n = 12$; what is the probability that the birthday of the twelve people will fall in twelve different months?

3.90 ¶. What is the probability that a random number between 1 and n divides n ?

3.91 ¶. Though Robin Hood is a wonderful archer (he hits the mark 9 times out of 10), he faces a difficult challenge in this tournament. In order to win, he must hit the center of the mark at least 4 times with the next 5 arrows. On the other hand, if he hit the mark 5 times out of 5, the county sheriff would recognize him. Let us suppose he can miss the mark at will: what is the probability of his winning the tournament?

3.92 ¶. A drug smuggler mixes drug pills with vitamin pills, hoping customs officers won't find him out. Of a total of 400 pills, only 5% are illegal ones. If the officers check 5 pills, what is their probability of finding an illegal one?

3.93 ¶. A box contains 90 balls numbered from 1 to 90. We sample without replacement 5 balls. What is the probability that they contain the balls 1, 2 and 3? Suppose we add three more balls numbered 1, 2 and 3 to the original 90 balls. What is now the probability that after producing a sample of size 5 the trick is discovered?

3.94 ¶. Let $n \geq k \geq r \geq 0$ be natural numbers and let X be a set of cardinality n , $|X| = n$. Define

$$\mathcal{P}_{k,r}(X) := \{(A, B) \mid B \subset A \subset X, |B| = r, |A| = k\}.$$

Show that

$$|\mathcal{P}_{k,r}| = \binom{n}{k} \binom{k}{r}.$$

3.95 ¶. Show that the number of strings of characters of k letters from an alphabet of n letters is n^k .

Show that the number of strings of characters with n letters from an alphabet $A = \{A_1, A_2, \dots, A_r\}$ where the letter A_i appears k_i times with $k_i \geq 0$ and $\sum_{i=1}^n k_i = n$ is

$$\frac{n!}{k_1! k_2! \cdots k_r!}.$$

3.96 ¶. Show that

$$\begin{aligned}
\binom{-1}{k} &= (-1)^k, & \binom{-2}{k} &= (-1)^k(k+1), \\
\binom{-n}{k} &= (-1)^k \binom{n+k-1}{k}, & \binom{2n}{n} 2^{-2n} &= (-1)^n \binom{-1/2}{n}, \\
\sum_{j=0}^k \binom{n}{k} n - k k - j t^j &= \binom{n}{k} (1+t)^k, & \frac{1}{n} \binom{2n-2}{n-1} 2^{-2n+1} &= (-1)^{n-1} \binom{1/2}{n}, \\
\sum_{j=0}^n \binom{a}{j} \binom{b}{n-j} &= \binom{a+b}{n}, & \sum_{j=0}^n \binom{n}{j}^2 &= \binom{2n}{n}, \\
\sum_{j=0}^n \frac{(2n)!}{j!^2 (n-j)!^2} &= \binom{2n}{n}^2, & \sum_{k=0}^n (C_k^n)^2 &= C_k^{2n}.
\end{aligned}$$

3.97 ¶. Let $|X| = n$ and let $\mathcal{P}_k(X) := \{A \subset X \mid |A| = k\} \subset \mathcal{P}(X)$ be the set of k -subsets of X and $\Pi_k(X)$ the set of k -partitions of X , i.e., of subsets $\{A_1, \dots, A_k\}$ of X which are disjoint and such that $X = \cup_{i=1}^k A_i$. Show that

$$|\mathcal{I}(X, \{1, 2, \dots, k\})| = k! |\mathcal{P}_k(X)|, \quad |S(X, \{1, 2, \dots, k\})| = k! |\Pi_k(X)|.$$

Moreover show that

$$|\Pi_k(X \cup \{x_0\})| = k |\Pi_k(X)| + |\Pi_{k-1}(X)|.$$

[Hint: Notice that the k -partitions of $X \cup \{x_0\}$ divide into the ones for which $\{x_0\}$ is one of the sets and the ones where x_0 is properly contained in one of the k -subsets.]

3.98 ¶. N balls numbered from 1 to N are successively located in N cells numbered from 1 to N starting from the first. What is the probability that a ball is located in the cell with the same number? [Hint: Compute first the probability that k balls are located in the corresponding cells.]

3.99 ¶¶. Let E_1, \dots, E_n be finite sets such that the intersection of k of them has always the same power, i.e., for all $i_1 < i_2 < \dots < i_k$ we have $|E_{i_1} \cap E_{i_2} \cap \dots \cap E_{i_k}| = c(k)$. Show that

$$\left| \bigcup_{i=1}^n E_i \right| = \sum_{i=1}^n (-1)^{k+1} \binom{n}{k} c(k).$$

In particular show that, if \widehat{P}_n denotes the family of permutations of n objects without fixed points,

$$\widehat{P}_n = \{\sigma \in P_n \mid \sigma_i \neq i\}$$

we have

$$|\widehat{P}_n| = n! \sum_{k=0}^n (-1)^k \frac{1}{k!}.$$

[Hint: Write $P_n \setminus \widehat{P}_n = \cup_{i=1}^n E_i$ with $E_i = \{\sigma \in P_n \mid \sigma_i = i\}$.]

3.100 ¶¶ Graphs. Many problems, both theoretical as well as of practical interest, often translate into graph problems.

Definition. A (symmetric) graph with vertices V is a subset G of $V \times V$ such that if $(u, v) \in G$, then $(v, u) \in G$, and $(v, v) \notin G$ for all $v \in V$.

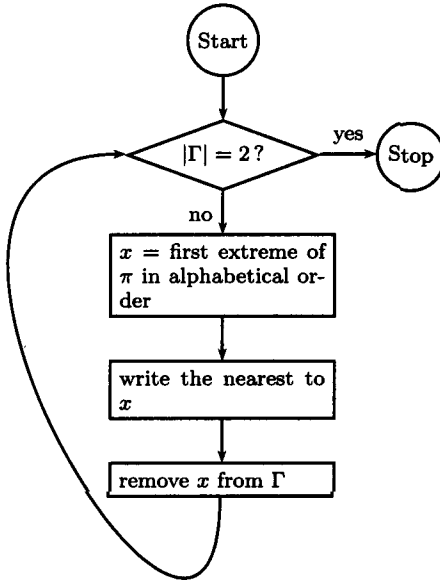


Figure 3.17. Passing from a tree Γ to a word π .

Two graphs (V, G) and (V', G') are said to be *isomorphic* if there is a bijection $\varphi : V \rightarrow V'$ such that $(u, v) \in G$ if and only if $(\varphi(u), \varphi(v)) \in G'$. Of course we can always decide if two finite graphs are isomorphic or not; however, the needed time can be very large in presence of many vertices: the best algorithms are just slightly more efficient than comparing the $n!$ bijections from V and V' .

To make the comparison more efficient, it is convenient to look at *invariants*. One such invariant is the number of connected components of a graph.

Definition. The connected component of v in G is the set

$$[v] := \left\{ w \in V \mid \exists u_0, u_1, \dots, u_r \in G \right. \\ \left. \text{such that } u_0 = v, u_r = w, \text{ and } (u_{i-1}, u_i) \in G \forall i = 1, \dots, r \right\}.$$

The number of connected components of G , $c(G)$, is clearly the same for isomorphic graphs. In particular, G is said to be *connected* if it has only one connected component: being connected is an invariant.

Another invariant is the *chromatic polynomial* introduced by George Birkhoff (1884–1944) in 1912.

Definition. A coloring of a graph (V, G) with $x \in \mathbb{N}$ colors is a mapping $\chi : V \rightarrow \{1, 2, \dots, x\}$ such that $\chi(u) \neq \chi(v)$ whenever $(u, v) \in G$, that is, such that adjacent vertices are colored differently.

The least number of distinct colors needed to color a graph is called the *chromatic number*, $\gamma(n)$, of the graph. Given a graph with n vertices and $x \geq \gamma(n)$, show that the number of coloring of G with x colors is given by a polynomial in x of degree $n = |V|$, called the *chromatic polynomial* of G .

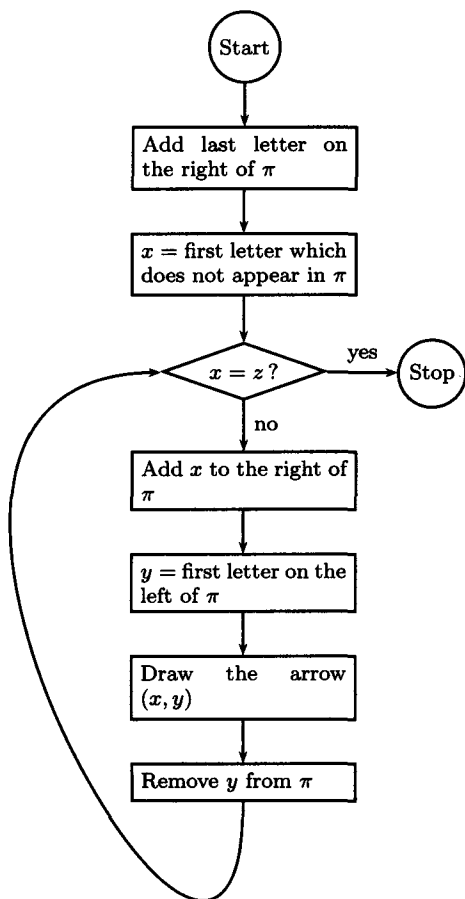


Figure 3.18. Passing from a word π to a tree Γ .

$$p_G(x) := \sum_{\Gamma \subset G} (-1)^{|\Gamma|} x^{c(\Gamma)}$$

where the sum is taken on all subgraphs Γ of G (including \emptyset and G).⁵ [Hint: Compute the number of wrong colorings.]

3.101 ¶¶ Trees. A *tree* Γ is a connected graph without cycles, a cycle being a sequence of distinct vertices u_1, \dots, u_n , $n \geq 2$, with $(u_k, u_{k+1}) \in G$ and $(u_n, u_0) \in G$.

Theorem (Cayley). The number of trees with n vertices is the number of words with $n - 2$ letters from an alphabet of n letters, i.e., n^{n-2} .

Figures 3.17 and 3.18 show how to construct a word from a tree Γ and a tree from a word.

⁵ There exist efficient algorithms to compute the chromatic polynomial of a graph.

3.102 ¶ The Pigeonhole Principle. Prove

Proposition. *Let X and Y be nonempty finite sets and let $\varphi : X \rightarrow Y$. There exists $y \in Y$ such that the fiber $\varphi^{-1}(y)$ contains at least $|X|/|Y|$ elements.*

Despite its simplicity, it is one of the most powerful methods of combinatorics. However, it is not always easy to understand how to use it. As an example of its applications we state, without proof, the following theorem.

Theorem (Erdős–Szekeres). *Let $a, b \in \mathbb{N}$, $n := ab + 1$ and let x_1, x_2, \dots, x_n be any n -sample of real numbers. Then the sequence contains either an increasing sequence of $a + 1$ numbers or a decreasing sequence of $b + 1$ numbers.*

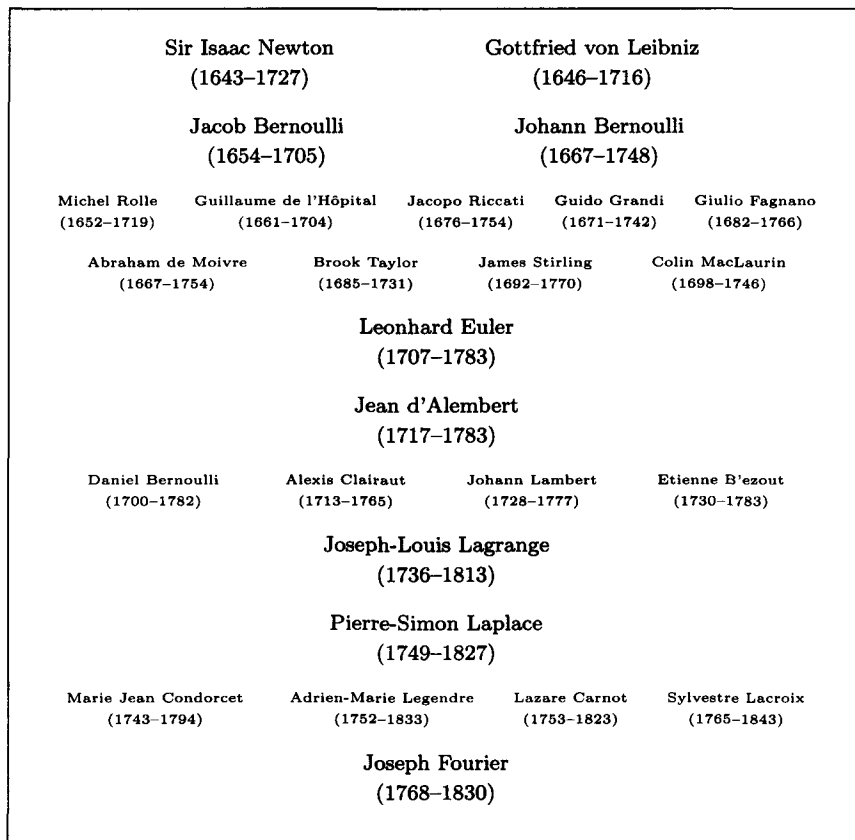


Figure 3.19. Infinitesimal analysis: a chronology from Newton and Leibniz to Fourier.

4. Complex Numbers

As already stated, the process of formation of numerical systems has been very slow. For instance, while Heron of Alexandria (IAD) and Archimedes of Syracuse (287BC–212BC) essentially accepted irrational numbers, working with their approximations, Diophantus of Alexandria (200–284) thought that equations with no integer solutions were not solvable; and only in the fifteenth century were negative numbers accepted as solutions of algebraic equations.¹ In the sixteenth century complex numbers enter the scene, with Girolamo Cardano (1501–1576) and Rafael Bombelli (1526–1573), in the resolution of algebraic equations as *surdes* numbers, that is numbers which are convenient to use in order to achieve correct real number solutions. But René Descartes (1596–1650) rejected complex roots and coined the term *imaginary* for these numbers. Despite the fact that complex numbers were fruitfully used by Jacob Bernoulli (1654–1705) and Leonhard Euler (1707–1783) to integrate rational functions and that several complex functions had been introduced, such as the complex logarithm by Leonhard Euler (1707–1783), complex numbers were accepted only after Carl Friedrich Gauss (1777–1855) gave a convincing geometric interpretation of them and proved the *fundamental theorem of algebra* (following previous researches by Leonhard Euler (1707–1783), Jean d’Alembert (1717–1783) and Joseph-Louis Lagrange (1736–1813)). Finally, in 1837 William R. Hamilton (1805–1865) introduced a formal definition of the system of complex numbers, which is essentially the one in use, giving up the mysterious imaginary unit $\sqrt{-1}$. Meanwhile complex functions reveal their importance in treating the equations of hydrodynamics and electromagnetism, and, in the eighteenth century develop into the *theory of functions of complex variables* with Augustin-Louis Cauchy (1789–1857), Karl Weierstrass (1815–1897) and G. F. Bernhard Riemann (1826–1866).

¹ For example, Antoine Arnauld (1612–1694) questioned that $-1 : 1 = 1 : -1$ by asking how a smaller could be to a greater as a greater to a smaller.



Figure 4.1. Girolamo Cardano (1501–1576) and Niccolò Fontana (1500–1557), called Tartaglia.

4.1 Complex Numbers

The development of the notion of complex numbers goes through their use in algebraic and differential problems and the understanding of their geometric properties. An a posteriori motivation is that they allow the solution of algebraic equations, as for instance $x^2 + 1 = 0$, that is not solvable in \mathbb{R} .

a. The system of complex numbers

4.1 Gauss plane. The set of complex numbers, denoted \mathbb{C} , is the *Gauss plane*, that is the Cartesian plane \mathbb{R}^2 with the operations of sum,

$$(a, b) + (c, d) := (a + c, b + d),$$

that is, the usual rule of summing vectors in \mathbb{R}^2 , and of *product*, defined by

$$(a, b) \cdot (c, d) := (ac - bd, ad + bc).$$

If we identify the axis of abscisses with \mathbb{R} in such a way that $(0, 0) = 0$ and $(1, 0) = 1$, and we introduce the *imaginary unit* i to indicate the vector $(0, 1)$, we see, on account of the computation rules previously defined, that $i^2 = i \cdot i = -1$, that $z = (a, b) = a(1, 0) + b(0, 1)$ is written as $z = a + ib$, and that the product of two complex numbers is written as

$$(a + ib)(c + id) = ac + iad + ibc + i^2bd = (ac - bd) + i(ad + bc).$$

It is easily seen that the properties (A) , (M) and (AM) of the real system \mathbb{R} , relative to the sum and the product, continue to hold in \mathbb{C} , 0 and 1 being this time respectively $0 := 0 + i0$, $1 := 1 + i0$. The inverse $1/z$ of the complex number $z = x + iy \neq 0$ is then given by

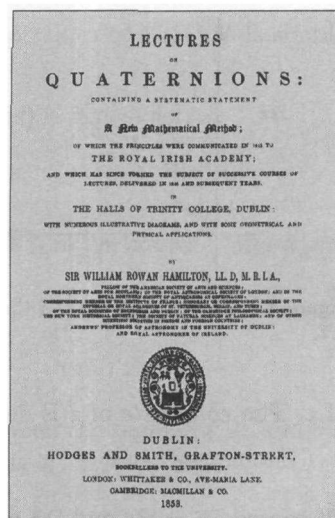


Figure 4.2. The frontispiece of the *Lectures on Quaternions* by William R. Hamilton (1805–1865).

$$\frac{1}{x + iy} = \frac{x - iy}{(x + iy)(x - iy)} = \frac{x - iy}{x^2 + y^2} = \frac{x}{x^2 + y^2} - i \frac{y}{x^2 + y^2}.$$

Consequently we can summarize saying that \mathbb{C} is a commutative field. Moreover, since the sum and product of complex numbers reduce to the sum and product of real numbers on the real axis, we can state that $\mathbb{R} \simeq \{x + iy \in \mathbb{C} \mid y = 0\}$ is a subfield of \mathbb{C} .

Of course there are several ways of ordering complex numbers; for instance, we can order them *lexicographically*: $(a, b) \prec (c, d)$ if $a < c$ or $a = c$ and $b < d$. However, none of all possible orderings is compatible with the field structure of \mathbb{C} and the order of \mathbb{R} . Otherwise, we would have either $i > 0$ or $i < 0$, as $i \neq 0$ being $0 \in \mathbb{R}$ and $i \notin \mathbb{R}$, thus, in both cases $-1 = i^2 > 0$: a contradiction. For this reason inequalities between complex numbers are meaningless.

4.2 Conjugation. Let $z := a + ib \in \mathbb{C}$. The numbers a and b , denoted also $a =: \Re z$ and $b =: \Im z$, are called the *real part* and the *imaginary part*

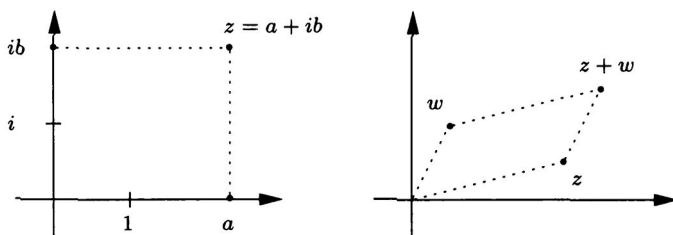


Figure 4.3. The sum of complex numbers.

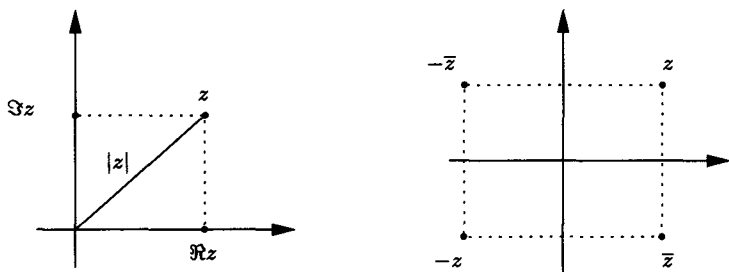


Figure 4.4. (a) $\Re z$, $\Im z$, $|z|$ and the argument θ of z . (b) Relative locations of $\pm z$ and $\pm \bar{z}$.

of z . The *conjugate* of z is defined by

$$\bar{z} := a - ib = \Re z - i\Im z.$$

Of course $\Re \bar{z} = \Re z$ and $\Im \bar{z} = -\Im z$. Consequently \bar{z} is the symmetric of z with respect to the real axis. The symmetric point of z with respect to the imaginary axis is $-\bar{z}$, and the symmetric of z with respect to the origin is $-z$. Moreover,

$$\Re z = \frac{z + \bar{z}}{2}, \quad \Im z = \frac{z - \bar{z}}{2i}.$$

4.3 ¶. Show that

$$\begin{aligned} \bar{\bar{z}} &= z, & \overline{z + w} &= \bar{z} + \bar{w}, & \overline{z \cdot w} &= \bar{z} \cdot \bar{w}, \\ \overline{\left(\frac{1}{w}\right)} &= \frac{1}{\bar{w}}, & \overline{\left(\frac{z}{w}\right)} &= \frac{\bar{z}}{\bar{w}} & \text{ if } w \neq 0. \end{aligned}$$

4.4 Absolute value or modulus. Let $z = a + ib \in \mathbb{C}$. Its *absolute value*, or *modulus*, is defined as the nonnegative real number

$$|z| := \sqrt{a^2 + b^2} = \sqrt{(\Re z)^2 + (\Im z)^2}.$$

Clearly $|z|$ is the Euclidean length of z in the Gauss plane (i.e., the distance between z and the origin) and agrees with the modulus in \mathbb{R} if z is real. Clearly

- (i) $|z| \geq 0$, $|z| = 0$ if and only if $z = 0$,
- (ii) **TRIANGLE INEQUALITY.** $|z + w| \leq |z| + |w|$.

4.5 ¶. Show that for all z and $w \in \mathbb{C}$ the following hold:

$$\begin{aligned} |z|^2 &= z\bar{z}, & |zw| &= |z||w|, & |\bar{z}| &= |z|, \\ |\Re z| &\leq |z|, & |\Im z| &\leq |z|, & ||z| - |w|| &\leq |z - w|, \\ \frac{1}{z} &= \frac{\bar{z}}{|z|^2} & \text{ if } z \neq 0. \end{aligned}$$

4.6 Hermitian product. The Hermitian product of z and w is simply $z\bar{w}$. If $w = a + ib$ and $z = c + id$, then

$$z\bar{w} = (c + id)(a - ib) = (ac + bd) + i(ad - bc)$$

from which we easily infer

- $z\bar{z} = x^2 + y^2 = |z|^2$,
- $\Re(z\bar{w})$ is the scalar product of z and w in \mathbb{R}^2 , in particular z and w are perpendicular if and only if $\Re(z\bar{w}) = 0$,
- the area of the triangle T with vertices 0 , z and w is

$$\text{Area}(T) = \frac{1}{2}|ad - bc| = \frac{1}{2}|\Im(z\bar{w})|.$$

For the last claim, denote by φ the angle between the segments $0z$ and $0w$ at the origin, and recall that $ac + bd = |z||w| \cos \varphi$ and that $\text{Area}(T) = |z||w| |\sin \varphi|$. Therefore

$$\begin{aligned} \text{Area}(T)^2 &= |z|^2 |w|^2 (1 - \cos^2 \varphi) \\ &= (c^2 + d^2)(a^2 + b^2) - (ac + bd)^2 = \cdots = (ad - bc)^2. \end{aligned} \quad (4.1)$$

4.7 Polar form of complex numbers: Argand's plane. A complex number $z = a + ib \in \mathbb{C}$, $z \neq 0$, can be represented in polar coordinates (r, θ) with center at the origin, r being the modulus of z and θ the angle between the real positive axis and the half-line from the origin through z , "measured counterclockwise and in radians," that is,

$$a = |z| \cos \theta, \quad b = |z| \sin \theta \quad (4.2)$$

i.e.,

$$z = |z|(\cos \theta + i \sin \theta). \quad (4.3)$$

Notice that the previous equality holds for all $z \in \mathbb{C}$. It is the *polar representation* of z . The number θ is called the *argument* or *phase* of z . Clearly, θ is determined by z up to an integer multiple of 2π , in particular $\theta(1) = 0$ modulo 2π . The argument of z , denoted by $\text{Arg}(z)$, must be understood as a *multivalued* function and not as a real-valued function. If we insist in considering the argument of z as a function from $\mathbb{C} \setminus \{0\}$ to \mathbb{R} , we have to choose a *determination*: that is, an interval $[a, a + 2\pi[$ where a unique value of the angle must be read. The restriction of the argument to this interval, called a *determination of the argument*, is denoted by $\arg^{(a)}(z)$. Among all determinations, two standard choices are $a = 0$, that is $\theta \in [0, 2\pi[$, often called the *principal determination*, commonly denoted by $\arg z$, and $a = -\pi$, that is $\theta \in [-\pi, \pi[$.

However, choosing a determination has drawbacks: first we have a discontinuity of the argument function $\arg^{(a)}(z)$ along the half-line through the origin that forms an angle a with the positive x -axis, where a jump of 2π between the values of the two sides of the half-line exists; secondly, addition formulas just do not hold for a determination, we in fact have

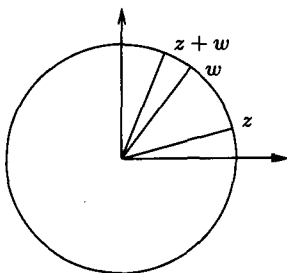


Figure 4.5. Multiplication of complex numbers.

$$\arg^{(a)}(z_1 z_2) = \arg^{(a)} z_1 + \arg^{(a)} z_2 - c$$

where

$$c = a + \begin{cases} 0 & \text{if } 2a \leq \arg^{(a)} z_1 + \arg^{(a)} z_2 < 2a + 2\pi, \\ 2\pi & \text{if } \arg^{(a)} z_1 + \arg^{(a)} z_2 \geq 2a + 2\pi. \end{cases}$$

4.8 ¶. Show that

$$\theta(z) = \begin{cases} \arctan \frac{y}{x} & \text{if } x > 0, \\ \frac{\pi}{2} & \text{if } x = 0 \text{ and } y > 0, \\ \arctan \frac{y}{x} + \pi & \text{if } x < 0 \text{ and } y > 0, \\ \arctan \frac{y}{x} - \pi & \text{if } x < 0 \text{ and } y \leq 0, \\ -\frac{\pi}{2} & \text{if } x = 0 \text{ and } y < 0, \end{cases}$$

where $z = x + iy$, is the determination of the argument on $[-\pi, \pi]$.

4.9 Multiplication in polar coordinates. If $z = \rho(\cos \theta + i \sin \theta)$ and $w = r(\cos \eta + i \sin \eta)$, on account of the addition formulas for the trigonometric functions and of the rule of multiplication for complex numbers, we get

$$\begin{aligned} zw &= \rho r[(\cos \theta \cos \eta - \sin \theta \sin \eta) + i(\cos \theta \sin \eta + \sin \theta \cos \eta)] \\ &= \rho r[\cos(\theta + \eta) + i \sin(\theta + \eta)]. \end{aligned} \quad (4.4)$$

That is, the modulus of the product is the product of moduli, while the argument of the product is the sum of the arguments of the factors. Geometrically, multiplying a vector $z \in \mathbb{C}$ by $w := |w|(\cos \eta + i \sin \eta)$ means dilating the vector by a factor $|w|$ and rotating it anticlockwise through an angle η . For instance iz is the anticlockwise rotation of z by 90 degrees. Thus dilations and rotations for plane geometry can all be expressed by complex multiplication, a useful fact in plane geometry (see, for example, Section 4.3.2).

4.10 de Moivre's formula. A trivial consequence of (4.4) is that

$$z^2 = \rho^2(\cos 2\theta + i \sin 2\theta)$$

if $z = \rho(\cos \theta + i \sin \theta)$, and, proceeding by induction, the following formula, *de Moivre's formula*, holds for every $n \in \mathbb{N}$,

$$z^n = \rho^n(\cos n\theta + i \sin n\theta). \quad (4.5)$$

4.11 Complex exponential. Set $f(\theta) := \cos \theta + i \sin \theta$, $\theta \in \mathbb{R}$. The multiplication rule yields the formula

$$f(\theta_1)f(\theta_2) = f(\theta_1 + \theta_2) \quad \forall \theta_1, \theta_2 \in \mathbb{R},$$

which is analogous to $a^{x_1}a^{x_2} = a^{x_1+x_2}$. We then define the *complex exponential* as the map $e^z : \mathbb{C} \rightarrow \mathbb{C}$ also denoted by \exp , given by

$$e^z = \exp(z) := e^{\Re z}(\cos \Im z + i \sin \Im z), \quad (4.6)$$

e being Euler's number. It is readily seen that

$$\begin{aligned} e^z e^w &= e^{z+w} & \forall z, w \in \mathbb{C}, \\ |e^z| &= e^{\Re z}. \end{aligned}$$

Clearly, if z is real, the complex and real exponential (with base e) agree; the novelty is in the definition

$$e^{i\theta} = \cos \theta + i \sin \theta, \quad (4.7)$$

which allows the use of an exponential notation for the trigonometric functions $\sin \theta$ and $\cos \theta$. Notice the very famous *Euler's identity*

$$e^{i\pi} = -1.$$

Actually, observing that for all $\theta \in \mathbb{R}$ we have

$$e^{i\theta} = \cos \theta + i \sin \theta, \quad e^{-i\theta} = \cos \theta - i \sin \theta,$$

we easily infer the following *Euler's formulas*:

$$\cos \theta = \frac{e^{i\theta} + e^{-i\theta}}{2}, \quad \sin \theta = \frac{e^{i\theta} - e^{-i\theta}}{2i}.$$

Finally, observe that (4.7) allow us to write any complex number in the shorter polar form

$$z = |z| e^{i \arg^{(a)} z}, \quad \forall a \in \mathbb{R}, \forall z \in \mathbb{R}.$$

b. The n -th roots

4.12 Proposition. Let $w \in \mathbb{C}$, $w \neq 0$, $n \in \mathbb{N}$, $n \geq 1$. The equation $z^n = w$ has exactly n distinct roots z_0, z_1, \dots, z_{n-1} given by

$$z_k := |w|^{1/n} \exp\left(i \frac{\arg(w) + 2k\pi}{n}\right) \quad k = 0, 1, \dots, n-1. \quad (4.8)$$

Proof. If z is a root of $z^n = w$, then

$$|z|^n = |w| \quad \text{and} \quad \text{Arg } z^n = \text{Arg } w.$$

The first equality yields $|z| = |w|^{1/n}$, while the second yields

$$\text{Arg } z = \frac{\text{Arg } w + 2k\pi}{n}, \quad k \in \mathbb{Z},$$

as $\text{Arg } z^n = n \text{Arg } z$. Therefore

$$z = |w|^{1/n} \exp\left(i \frac{\arg(w) + 2k\pi}{n}\right), \quad \forall k \in \mathbb{Z}.$$

Since the values $(\arg(w) + 2k\pi)/n$ repeat periodically with period $2\pi/n$, the only distinct values in $[0, 2\pi[$ correspond to $k = 0, 1, \dots, n-1$. Thus we conclude that z ought to be one of the z_k 's. Finally, one checks that all the z_k 's are solutions of $z^n = w$. \square

The n distinct solutions z_0, z_1, \dots, z_{n-1} of the equation $z^n = w$ in (4.8) are called the *complex* or *algebraic* n -th roots of w .

Proposition 4.12 applies also to $w = a \in \mathbb{R}$. If $a > 0$, we have $a = |a| = |a| \exp(i \cdot 0)$, thus

$$\sqrt[n]{a} = |a|^{1/n} \exp\left(i \frac{2\pi k}{n}\right) \quad k = 0, 1, \dots, n-1.$$

If $a < 0$, we have $a = |a| \exp(i\pi)$, hence

$$\sqrt[n]{a} = |a|^{1/n} \exp\left(i \frac{(2k+1)\pi}{n}\right) \quad k = 0, 1, \dots, n-1.$$

In particular, for $a > 0$, we rediscover the arithmetic root $a^{1/n}$ corresponding to $k = 0$, and, in case n is even, for $k = n/2$ we also find

$$\sqrt[n]{a} = a^{1/n} (\cos \pi + i \sin \pi) = -a^{1/n},$$

as n -th root of a , that is the negative real n -th root of a . If $a < 0$ and n is odd, $n = 2h + 1$, then for $k = h$ $\sqrt[n]{a} = |a|^{1/n} (\cos \pi + i \sin \pi) = -|a|^{1/n}$ is one of the complex roots and is the only real n -th root of a .

Notice that for every $k = 0, \dots, n-1$ the argument of z_k is the argument of z_{k-1} plus $2\pi/n$. Consequently, the n -th roots of a complex number w represent a regular n -sided polygon, inscribed in a circle at 0, with radius $|w|^{1/n}$ and one vertex at $|w|^{1/n} e^{i \arg(w)/n}$.

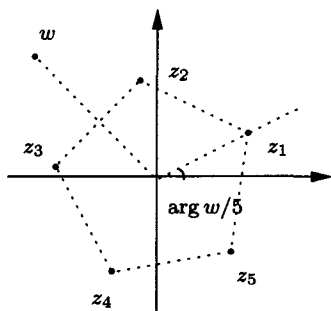


Figure 4.6. The 5th roots of a complex number.

4.13 Roots of unity. In particular Proposition 4.12 yields that the n solutions of $z^n = 1$ are given by

$$\omega_{n,k} := \exp\left(i\frac{2\pi k}{n}\right), \quad k = 0, \dots, n-1.$$

Let $\omega := \omega_{n,1} = e^{i2\pi/n}$. Then the n -th roots of unity are the numbers

$$1 = \omega^0, \omega, \omega^2, \omega^3, \dots, \omega^{n-1}. \quad (4.9)$$

Notice that obviously $\omega^n = 1$. Moreover, comparing with (4.8), if $z_1 := |w|^{1/n} \exp(i \arg(w)/n)$, then the n -th roots of w can be written as

$$z_1, z_1\omega, z_1\omega^2, \dots, z_1\omega^{n-1}. \quad (4.10)$$

c. Complex exponential and logarithm

The exponential function $z \rightarrow e^z$ is actually a map from \mathbb{C} into \mathbb{C} . Observe that $e^z \neq 0$ everywhere since $|e^z| = e^{\Re z} \neq 0$.

4.14 Proposition. *The complex exponential is periodic with period $2\pi i$, i.e.,*

$$\exp(z + 2\pi i) = \exp(z) \quad \forall z \in \mathbb{C}.$$

Moreover for any $a \in \mathbb{R}$, the restriction of the exponential map e^z to

$$I_a := \left\{ z \in \mathbb{C} \mid a \leq \Im z < a + \pi \right\}$$

is a bijective map onto $\mathbb{C} \setminus \{0\}$.

Proof. Formula (4.7) yields

$$\exp(i(y + 2k\pi)) = \exp(iy) \quad \forall y \in \mathbb{R},$$

that is, e^z is $2\pi i$ -periodic. Fix $a \in \mathbb{R}$ and assume that $e^{z_1} = e^{z_2}$, i.e., $e^{z_1 - z_2} = 1$. Then $z_1 - z_2 = 2k\pi i$, from which we infer $z_1 = z_2$, since $z_1, z_2 \in I_a$. \square

Given $z \in \mathbb{C}$, $z \neq 0$, every $w \in \mathbb{C}$ such that $e^w = z$ is called a *natural logarithm* of z . More precisely,

4.15 Definition. For any $a \in \mathbb{R}$, The inverse function of the restriction of the complex exponential e^z to

$$I_a := \{z \mid a \leq \Im z < a + 2\pi\}$$

is called a determination of the complex logarithm,

$$\log^{(a)} : \mathbb{C} \setminus \{0\} \rightarrow I_a \subset \mathbb{C}.$$

When $a = -\pi$, we denote $\log^{(-\pi)} w$ by $\log w$ and call it the principal logarithm.

By definition

$$e^{\log^{(a)} w} = w \quad \forall w \in \mathbb{C} \setminus \{0\}$$

and

$$\log^{(a)}(e^z) = z \quad \text{if and only if} \quad z \in I_a.$$

4.16 Proposition. For any $a \in \mathbb{R}$ and $w \in \mathbb{C} \setminus \{0\}$ we have

$$\log^{(a)} w = \log |w| + i \arg^{(a)} w. \quad (4.11)$$

Proof. Let $z : x + iy = \log^{(a)} w$. Then $w = e^z$ and $z \in I_a$ if and only if

$$\begin{cases} w = e^x e^{iy}, \\ a \leq y < a + 2\pi, \end{cases} \quad \text{if and only if} \quad \begin{cases} |w| = e^x, \\ e^{iy} = \frac{w}{|w|}, a \leq y < a + 2\pi, \end{cases}$$

from which we infer $x = \log |w|$ and $y = \arg^{(a)} w = \arg^{(a)} z$. \square

4.17 Example. Since $i = \cos \frac{\pi}{2} + i \sin \frac{\pi}{2}$, i.e., $\arg i = \frac{\pi}{2}$, we have $\log i = i \frac{\pi}{2}$ or $i^i = e^{-\pi/2}$.

On account of Proposition 4.16, all determinations of the logarithm are discontinuous as the corresponding determination of the argument. In particular $\log^{(a)} w$ is singular along the half-line that has an angle a with the positive x -axis, with a jump of $2\pi i$ along this half-line. Also, some care is necessary to compute with logarithms since the argument of a product is not in general the sum of the arguments. In fact, if $z, w \in \mathbb{C} \setminus \{0\}$, we have

$$\log^a(zw) = \log^{(a)} z + \log^{(a)} w - ic$$

where

$$c = a + \begin{cases} 0 & \text{if } \arg^{(a)}(z) + \arg^{(a)}(w) < 2a + 2\pi, \\ 2\pi & \text{if } \arg^{(a)}(z) + \arg^{(a)}(w) \geq 2a + 2\pi. \end{cases}$$

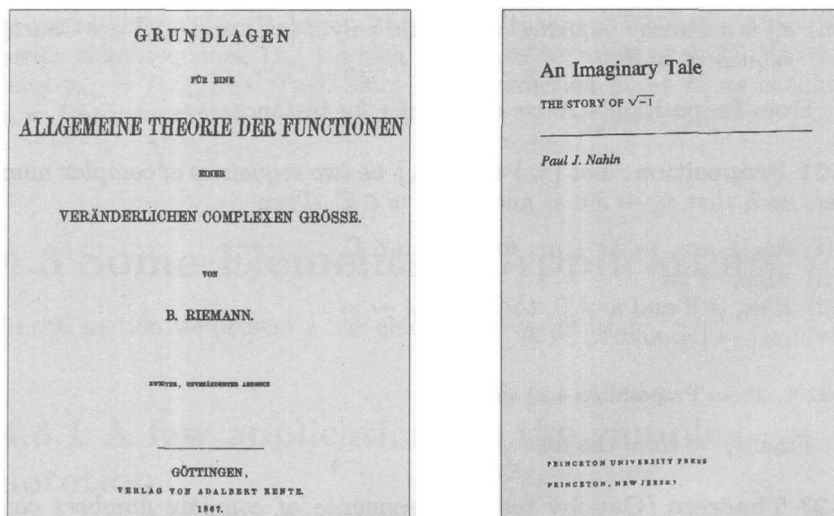


Figure 4.7. The frontispieces of a treatise on functions of complex variables by G. F. Bernhard Riemann (1826–1866) and of a popular book about $\sqrt{-1}$.

4.2 Sequences of Complex Numbers

a. Definitions

The limit of a sequence of complex numbers is defined similarly to the real case.

4.18 Definition. Let $\{z_n\} \subset \mathbb{C}$ be a sequence of complex numbers. We say that $\{z_n\}$ converges to the complex number z_0 if $|z_n - z_0| \rightarrow 0$, that is

$$\forall \epsilon > 0 \exists \bar{n} \text{ such that } |z_n - z_0| < \epsilon \quad \forall n \geq \bar{n}.$$

We say that $\{z_n\} \subset \mathbb{C}$ diverges if $|z_n| \rightarrow +\infty$ as a sequence of real numbers.

As in the real case

4.19 Definition. We say that a sequence $\{z_n\} \subset \mathbb{C}$ is a Cauchy sequence if $\forall \epsilon > 0$ there is \bar{n} such that $|z_n - z_m| < \epsilon$ for all $n, m \geq \bar{n}$.

From the inequalities

$$|x|, |y| \leq |z| = \sqrt{x^2 + y^2} \leq |x| + |y| \quad (4.12)$$

for all $z = x + iy \in \mathbb{C}$, we easily infer

4.20 Proposition. Let $\{z_n\}$, $z_n := x_n + iy_n$, be a sequence of complex numbers, and let $z_0 := x_0 + iy_0 \in \mathbb{C}$. Then

- (i) $z_n \rightarrow z_0 \in \mathbb{C}$ if and only if $x_n \rightarrow x_0$ and $y_n \rightarrow y_0$.

- (ii) z_n is a Cauchy sequence in \mathbb{C} if and only if $\{x_n\}$ and $\{y_n\}$ are Cauchy sequences in \mathbb{R} .

From Proposition 4.20 we easily infer for instance

4.21 Proposition. Let $\{z_n\}$ and $\{w_n\}$ be two sequences of complex numbers such that $z_n \rightarrow z \in \mathbb{C}$ and $w_n \rightarrow w \in \mathbb{C}$. Then

- (i) $\lambda z_n + \mu w_n \rightarrow \lambda z + \mu w$ for all $\lambda, \mu \in \mathbb{C}$,
- (ii) $z_n w_n \rightarrow zw$,
- (iii) if $w_n \neq 0$ and $w \neq 0$, then $z_n/w_n \rightarrow z/w$,
- (iv) $|z_n| \rightarrow |z|$, and $\bar{z}_n \rightarrow \bar{z}$.

4.22 ¶. Prove Propositions 4.20 and 4.21.

Finally, we have the following.

4.23 Theorem (Cauchy test). A sequence of complex numbers converges if and only if it is a Cauchy sequence.

This follows again from Proposition 4.20 if we take into account the Cauchy test for real sequences. For the same reason the Bolzano–Weierstrass theorem, Theorem 2.43 extends to complex sequences

4.24 Theorem (Bolzano–Weierstrass). Any bounded sequence of complex numbers has a convergent subsequence.

b. Weierstrass's theorem

At this point we could introduce the notions of *limit* and of *continuity* for functions of a complex variable and develop a theory similar to the real case. Instead, we prefer to postpone this study in the context of *metric spaces*. Here we confine ourselves to defining continuity for functions $f : \mathbb{C} \rightarrow \mathbb{R}$. A function $f : \mathbb{C} \rightarrow \mathbb{R}$ is said to be continuous at z_0 if for every sequence $\{z_n\}$ converging to z_0 we have $f(z_n) \rightarrow f(z_0)$; and f is *continuous* if it is continuous at each $z_0 \in \mathbb{C}$. Finally we say that $f(z) \rightarrow +\infty$ as $|z| \rightarrow +\infty$ if $\forall M > 0$ there exists $r > 0$ such that $f(z) > M$ for all z such that $|z| > r$.

4.25 Theorem (Weierstrass). Let $f : \mathbb{C} \rightarrow \mathbb{R}$ be a continuous function such that $f(z) \rightarrow +\infty$ as $|z| \rightarrow \infty$. Then f attains its minimum at a point $z_0 \in \mathbb{C}$.

Proof. Let $L := \inf\{f(z) \mid z \in \mathbb{C}\}$. It suffices to show that $f(z_0) = L$. From the characterization of the infimum we infer $-\infty \leq L < +\infty$, the existence of a minimizing sequence $\{y_n\} \subset f(\mathbb{C}) \subset \mathbb{R}$ such that $y_n \rightarrow L$ and of a sequence $\{z_k\} \subset \mathbb{C}$ such that $f(z_k) = y_k$. The sequence $\{z_k\}$ is bounded, otherwise we could find a subsequence $\{z_{n_k}\}$ of $\{z_k\}$ with $|z_{n_k}| \rightarrow \infty$, hence $f(z_{n_k}) \rightarrow +\infty$. Since $y_{n_k} = f(z_{n_k}) \rightarrow L$, we would get $L = +\infty$:

a contradiction. The Bolzano–Weierstrass theorem, Theorem 4.25, then yields a subsequence $\{z_{n_k}\}$ which converges to some $z_0 \in \mathbb{C}$. We then have $y_{n_k} = f(z_{n_k}) \rightarrow f(z_0)$. Since by construction $y_n \rightarrow L$, we conclude $L = f(z_0)$, as wanted. \square

4.3 Some Elementary Applications

In this section we present a few elementary applications.

4.3.1 A few applications of the complex notation

When dealing with trigonometric formulas, but actually in many instances, the complex notation simplifies computations a great deal.

4.26 Uniform circular motion. Recall that the harmonic motion of a point $P(t)$ on the unit circle of \mathbb{R}^2 , that starts at $t = 0$ from $(1, 0)$ with angular velocity ω , is given by

$$\begin{cases} x(t) = \cos \omega t, \\ y(t) = \sin \omega t, \end{cases}$$

see, e.g., Proposition 6.25 of [GM1]. Thus, introducing complex notation, the uniform circular motion on the unit circle is described by the function $P : \mathbb{R} \rightarrow \mathbb{C}$ given by $P(t) = e^{i\omega t}$. This formula already appears as a great simplification of the description of the harmonic motion on the unit circle.

However the simplification that can be obtained using complex numbers and complex notation is even more evident if one notices that it is easier to compute with powers than with sine and cosine. For instance, if we define for $z(t) : \mathbb{R} \rightarrow \mathbb{C}$, $z(t) = x(t) + iy(t)$,

$$\begin{aligned} z'(t) &= Dz(t) := x'(t) + iy'(t), \\ \int_0^t z(s) ds &:= \int_0^t x(s) ds + i \int_0^t y(s) ds, \end{aligned}$$

then we have

$$D(e^{i\omega t}) = i\omega e^{i\omega t}, \quad \int_0^t e^{i\omega s} ds = \frac{e^{i\omega t} - 1}{i\omega}$$

for all $\omega \in \mathbb{R}$. Clearly these formulas are handier than $D(e^{at} \cos(bt)) = ae^{at} \cos(bt) - be^{at} \sin(bt)$, and $D(e^{at} \sin(bt)) = ae^{at} \sin(bt) + be^{at} \cos(bt)$, or the corresponding formulas for the primitives.

4.27 Example. Euler's formulas yield that the sine and cosine function are just a superposition of two uniform circular motions with opposite angular velocities of ± 1 .

4.28 Complex solutions of the oscillation equation. Consider the differential equation

$$a x''(t) + b x'(t) + c = 0$$

and look for solutions $x : \mathbb{R} \rightarrow \mathbb{C}$. The characteristic equation

$$a\lambda^2 + b\lambda + c = 0$$

has two roots λ_1 and λ_2 that are either distinct or equal. In the first case, $\lambda_1 \neq \lambda_2$, $e^{\lambda_1 t}$ and $e^{\lambda_2 t}$ are solutions and, on account of the principle of superposition

$$c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}, \quad c_1, c_2 \in \mathbb{C}$$

are solutions, too. In the second case, $\lambda_1 = \lambda_2 =: \lambda$, the functions $e^{\lambda t}$ and $t e^{\lambda t}$ are solutions as well as all functions of the type

$$(c_1 + c_2 t) e^{\lambda t}, \quad c_1, c_2 \in \mathbb{C}.$$

Exactly as in the real case (see, e.g., Section 6.1.3 of [GM1]), one can then conclude that, in fact, these are *all* solutions.

4.29 Prostapheresis formulas. In complex notation they are written as

$$\begin{cases} e^{i\alpha} + e^{i\beta} = 2 \cos \frac{\alpha-\beta}{2} e^{i\frac{\alpha+\beta}{2}}, \\ e^{i\alpha} - e^{i\beta} = 2i \sin \frac{\alpha-\beta}{2} e^{i\frac{\alpha+\beta}{2}} \end{cases} \quad (4.13)$$

and can be easily deduced. In fact, writing

$$\alpha = \frac{\alpha + \beta}{2} + \frac{\alpha - \beta}{2}, \quad \beta = \frac{\alpha + \beta}{2} - \frac{\alpha - \beta}{2}$$

it suffices to note that

$$\begin{cases} e^{i\alpha} + e^{i\beta} = e^{i\frac{\alpha+\beta}{2}} \{ e^{i\frac{\alpha-\beta}{2}} + e^{-i\frac{\alpha-\beta}{2}} \}, \\ e^{i\alpha} - e^{i\beta} = e^{i\frac{\alpha+\beta}{2}} \{ e^{i\frac{\alpha-\beta}{2}} - e^{-i\frac{\alpha-\beta}{2}} \}. \end{cases}$$

Therefore they are a trivial consequence of Euler's formulas.

4.30 Beating phenomenon. This is a phenomenon which occurs when we sum two sinusoids of slightly different pulses, compare, e.g., 6.13 of [GM1]. As a simple example, let us show that the same phenomenon appears when we sum sinusoidal signals with different amplitudes. If $f_1(t) = A_1 \cos(\omega_1 t + \varphi_1)$, $f_2(t) = A_2 \cos(\omega_2 t + \varphi_2)$, we have

$$f_1(t) = \Re(c_1 e^{i\omega_1 t}), \quad f_2(t) = \Re(c_2 e^{i\omega_2 t})$$

where $c_1 = A_1 e^{i\varphi_1}$ and $c_2 = A_2 e^{i\varphi_2}$. Hence

$$f_1(t) + f_2(t) = \Re(c_1 e^{i\omega_1 t} + c_2 e^{i\omega_2 t}).$$

Factoring out the mean oscillation $\frac{\omega_1 + \omega_2}{2}$ we then find

$$c_1 e^{i\omega_1 t} + c_2 e^{i\omega_2 t} = e^{i\frac{\omega_1 + \omega_2}{2}t} \left\{ c_1 e^{i\frac{\omega_1 - \omega_2}{2}t} + c_2 e^{-i\frac{\omega_1 - \omega_2}{2}t} \right\}. \quad (4.14)$$

The explicit computation of the real part is of course complicated, but, without performing such a computation, we see that we are in the presence of a signal with pulse $\frac{\omega_1 + \omega_2}{2}$ and amplitude varying periodically with pulse $\frac{|\omega_1 - \omega_2|}{2}$.

4.3.2 A few applications to elementary Euclidean geometry

Translations, dilations and rotations are the typical transformations of Euclidean geometry of the plane. As we have seen, after introducing an orthonormal reference frame, they have a natural algebraic counterpart in the operation of sum and product in the Gauss plane. Therefore it is not surprising that the use of complex numbers permits a particularly simple algebraization of the geometry in the plane.

4.31 Straight line through two points $a \neq b \in \mathbb{C}$. The point $z \in \mathbb{C}$ is in the line through a and b if and only if $z - a$ is a real multiple of $b - a$, equivalently if and only if $(z - a)/(b - a)$ is real, that is

$$\frac{z - a}{b - a} = \frac{\bar{z} - \bar{a}}{\bar{b} - \bar{a}}, \quad \text{or} \quad \Im\left(\frac{z - a}{b - a}\right) = 0.$$

Consequently the two open half-planes bounded by that line are described by

$$\left\{ z \mid \Im\left(\frac{z - a}{b - a}\right) < 0 \right\} \quad \text{and} \quad \left\{ z \mid \Im\left(\frac{z - a}{b - a}\right) > 0 \right\}.$$

4.32 Perpendicular lines. Let a, b, c be three distinct points in the plane. Since an anticlockwise rotation by 90 degrees translates into a multiplication by i , the lines through a and b and through a and c are perpendicular if and only if $c - a/b - a$ is purely imaginary, that is

$$\frac{c - a}{b - a} = -\frac{\bar{c} - \bar{a}}{\bar{b} - \bar{a}}$$

or

$$\Re((c - a)(\bar{b} - \bar{a})) = 0.$$

Consequently the line through a and perpendicular to the line through a and b is

$$\left\{ z \in \mathbb{C} \mid (\bar{b} - \bar{a})(z - a) + (b - a)(\bar{z} - \bar{a}) = 0 \right\}.$$

4.33 Similarity of two triangles. Let $s_1, s_2, s_3 \in \mathbb{C}$ and $t_1, t_2, t_3 \in \mathbb{C}$ be the vertices of two triangles S and T . We recall that S and T are similar (with the ordered vertices) if the dilation and rotation which moves $t_2 - t_1$ to $s_2 - s_1$ moves also $t_3 - t_1$ onto $s_3 - s_1$. Since rotations and dilations translate into a complex multiplication, S and T are similar if and only if, for some $b \in \mathbb{C}$ we have $w_2 - w_1 = b(z_2 - z_1)$, then also $w_3 - w_1 = b(z_3 - z_1)$, that is,

$$\frac{z_3 - z_1}{z_2 - z_1} = \frac{w_3 - w_1}{w_2 - w_1}.$$

a. Special points of a triangle

4.34 Circumcenter. The perpendicular bisectors to the three sides of an arbitrary triangle meet at a point. It is called the *circumcenter* of the triangle, and it is the center of the circumcircle of the triangle.

Let $a, b, c \in \mathbb{C}$ be the vertices. The middle points of the three sides are respectively $(a+b)/2$, $(a+c)/2$ and $(b+c)/2$. The equations of the bisectors are consequently

$$\begin{cases} (\bar{b} - \bar{c})z + (b - c)\bar{z} = |b|^2 - |c|^2, \\ (\bar{c} - \bar{a})z + (c - a)\bar{z} = |c|^2 - |a|^2, \\ (\bar{a} - \bar{b})z + (a - b)\bar{z} = |a|^2 - |b|^2, \end{cases}$$

that, solved in z , give for the circumcenter

$$o = \frac{|a|^2(b - c) + |b|^2(c - a) + |c|^2(a - b)}{\bar{a}(b - c) + \bar{b}(c - a) + \bar{c}(a - b)}.$$

4.35 Barycenter or centroid. The three medians (the lines connecting each vertex to the middle point of the opposite side) meet at a point. If a, b, c are the vertices, $(b+c)/2$, $(a+c)/2$ and $(a+b)/2$ are the midpoints of the corresponding opposite sides. The intersection point of two medians can be obtained solving in $\lambda, \mu \in \mathbb{R}$ the system

$$\begin{cases} z = \lambda a + (1 - \lambda) \frac{b+c}{2}, \\ z = \mu c + (1 - \mu) \frac{a+b}{2}. \end{cases}$$

Subtracting, we easily infer, since the triangle is nondegenerate, that the previous system has one solution given by $\lambda = \mu = 1/3$, thus concluding that the barycenter is

$$z = \frac{a + b + c}{3},$$

z belonging also to the third median. Notice that it is easy now to prove that the intersection of the medians is two thirds of the way along from each of their vertices.

4.36 Orthocenter. The altitudes of a triangle (that is the perpendiculars dropped down from vertices onto the opposite side) intersect at a single point: the *orthocenter*. Fix a reference with origin at the circumcenter o of the triangle, and, in this reference, let $a, b, c \in \mathbb{C}$ be the three vertices of the triangle, so that $|a| = |b| = |c|$. We claim that the point

$$p = a + b + c$$

is the orthocenter. In fact, since $p - a = b + c$, and $|b| = |c|$, $p - a$ is perpendicular to the side bc . By symmetry, p is also on the perpendicular from b to ca and from c to ab .

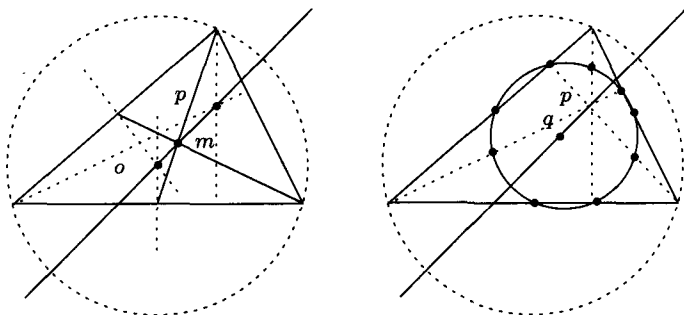


Figure 4.8. (a) Euler's line. (b) The nine-point circle.

4.37 ¶ Incenter. Show that the three angle bisectors of a triangle meet at a point called the *incenter*.

4.38 Euler's line. In any triangle, the circumcenter, the orthocenter and the barycenter lie on a straight line. In fact in a reference in which the circumcenter is the origin, the barycenter m and the orthocenter are respectively

$$m = \frac{1}{3}(a + b + c) \quad \text{and} \quad p = a + b + c,$$

by 4.35 and 4.36.

We also have the following theorem due to Karl Feuerbach (1800–1834), but probably already known to Charles Brianchon (1783–1864) and Jean-Victor Poncelet (1788–1867).

4.39 Theorem (The nine-point circle). Let $a, b, c \in \mathbb{C}$ be the vertices of a triangle that for convenience we think to be inscribed in a unitary circle, i.e., that is $|a| = |b| = |c| = 1$. Let q be the midpoint of the segment connecting the circumcenter and the orthocenter, that is, in the chosen frame,

$$q := \frac{a + b + c}{2}.$$

Then the circle with center q and radius $1/2$ goes through

- (i) the midpoints of the three sides,
- (ii) the midpoints of the segments joining the orthocenter with the three vertices,
- (iii) the feet of the three perpendiculars from the vertices to the opposite sides.

Proof. It suffices to show that each of those points has distance $1/2$ from q . The distance of the midpoint of bc from q is

$$\left| q - \frac{b + c}{2} \right| = \left| \frac{a}{2} \right| = \frac{|a|}{2} = \frac{1}{2}.$$

The midpoint of the segment joining the orthocenter p with a is $(a + (a + b + c))/2 = (a + 2q)/2$, hence

$$\left| q - \frac{a + 2q}{2} \right| = \left| \frac{a}{2} \right| = \frac{1}{2}.$$

Finally, one sees that the foot of the perpendicular from a to bc is

$$\eta = \left(\frac{a + b + c}{2} - \frac{bc}{2a} \right)$$

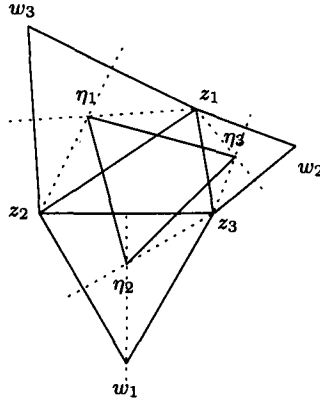


Figure 4.9. Napoleon's theorem.

hence

$$|q - \eta| = \left| \frac{bc}{2a} \right| = \frac{|b||c|}{2|a|} = \frac{1}{2}.$$

□

b. Equilateral triangles

4.40 Proposition. Let z_1, z_2, z_3 be the vertices of a triangle (listed anticlockwise) and let $\omega := \exp(i2\pi/3)$ be the second of the 3rd roots of unity. Then the triangle $z_1 z_2 z_3$ is equilateral if and only if

$$z_1 + \omega z_2 + \omega^2 z_3 = 0.$$

Proof. In fact, $z_1 z_2 z_3$ is equilateral if it is similar to the triangle of the 3rd roots of unity, $1, \omega, \omega^2$. Then, according to 4.33,

$$\frac{z_3 - z_1}{z_2 - z_1} = \frac{\omega^2 - 1}{\omega - 1} = \omega + 1,$$

i.e., $z_3 + \omega z_1 - (\omega + 1)z_2 = 0$. Since $\omega^2 + \omega + 1 = 0$, we infer

$$z_3 + \omega z_1 + \omega^2 z_2 = 0$$

and, multiplying by ω^2 , the conclusion. □

4.41 Napoleon's theorem. It is said that Napoleon stated and proved the following result. On each side of an arbitrary triangle draw the exterior equilateral triangle. Then the barycenters of these three equilateral triangles are the vertices of a fourth equilateral triangle. In fact, listing the vertices anticlockwise, if z_1, z_2, z_3 are the vertices of the triangle and $w_3 z_2 z_1, w_1 z_3 z_2, w_2 z_1 z_3$ are the exterior equilateral triangles, we have

$$\begin{cases} z_2 + \omega z_1 + \omega^2 w_3 = 0, \\ w_1 + \omega z_3 + \omega^2 z_2 = 0, \\ z_3 + \omega w_2 + \omega^2 z_1 = 0, \end{cases}$$

and the barycenters of the exterior triangles are given by

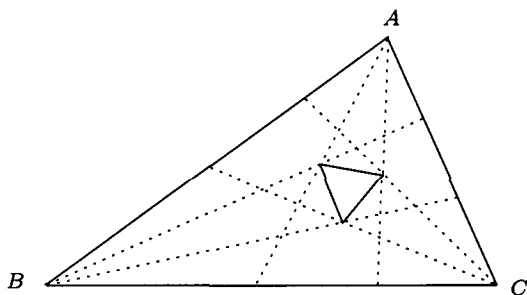


Figure 4.10. Morley's equilateral triangle.

$$\begin{cases} \eta_3 := \frac{z_1 + z_2 + w_3}{3}, \\ \eta_1 := \frac{z_2 + z_3 + w_1}{3}, \\ \eta_2 := \frac{z_1 + z_3 + w_2}{3}. \end{cases}$$

Therefore

$$\begin{aligned} \eta_1 + \omega\eta_2 + \omega^2\eta_3 &= \frac{1}{3}(w_1 + z_2 + z_3) + \frac{\omega}{3}(z_3 + w_2 + z_1) + \frac{\omega^2}{3}(z_2 + z_1 + w_3) \\ &= \frac{1}{3}((w_1 + \omega z_3 + \omega^2 z_2) + (z_3 + \omega w_2 + \omega^2 z_1) + (z_2 + \omega z_1 + \omega^2 w_3)) \\ &= 0. \end{aligned}$$

The following result, discovered by Frank Morley (1860–1937), is quite surprising

4.42 Theorem (Morley). *The intersections of the adjacent pairs of angle trisectors of an arbitrary triangle are the vertices of an equilateral triangle.*

4.43 Lemma. *Suppose that t_1, t_2, t_3, t_4 are points on the unit circle. Then the extensions of the chords joining the points t_1, t_2 and t_3, t_4 meet at*

$$z = \frac{\bar{t}_1 + \bar{t}_2 - \bar{t}_3 - \bar{t}_4}{\bar{t}_1 \bar{t}_2 - \bar{t}_3 \bar{t}_4}.$$

4.44 ¶. Prove Lemma 4.43.

Proof of Morley's theorem. For the sake of convenience assume that the triangle ABC is inscribed in the unit circle, $A = 1$, $\angle AOB = 3\gamma$, $\angle AOC = 3\beta$, $\beta < 0$, and $\angle BOC = 3\alpha$. Since circumferential angles are half the corresponding central angles, in order to find the intersections of the trisectors of vertex angles it suffices to trisect the corresponding central angles. We then call $B = c^3$ in such a way that the intersection of the trisectors of the angle in c with the circle are the points c and c^2 . Similarly we set $C = b^3$, $3b < 0$, in such a way that the corresponding intersections are b and b^2 . The arguments of the intersections of the trisectors of the angle in A are

$$\alpha + 3\gamma = -\beta + 2\gamma + \frac{2\pi}{3}, \quad 2\alpha + 3\gamma = -2\beta + \gamma + \frac{4\pi}{3},$$

consequently the intersections of the trisectors of the angle A with the circle are the points $\omega b^2 c$ and $\omega b c^2$ where $\omega := e^{i2\pi/3}$, see Figure 4.11.

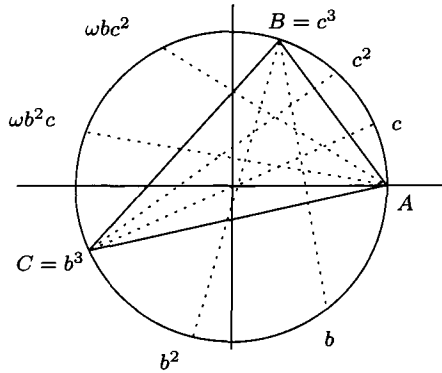


Figure 4.11. Intersections of the angle trisectors with the circumscribed circle.

If P, Q, R are the vertices of the triangle obtained intersecting adjacent trisectors, we compute, on account of Lemma 4.43,

$$\begin{aligned}
 p &= \frac{b^{-2} + c^{-3} - b^{-3} - c^{-2}}{b^{-2}c^{-3} - b^{-3}c^{-2}} = \frac{bc^3 + b^3 - c^3 - b^3c}{b - c} \\
 &= (b^2 + bc + c^2) - bc(b + c), \\
 q &= \frac{1 + b^{-2}c^{-1}\omega^{-2} - b^{-3} - c^{-1}}{b^{-2}c^{-1}\omega^{-2} - b^{-3}c^{-1}} = \frac{b^3c + b\omega - c - b^3}{b\omega - 1} \\
 &= \omega^2 \left(c(b^2 + b\omega^2 + \omega) - b(b + \omega^2) \right), \\
 r &= \frac{1 + b^{-1}c^{-2}\omega^{-1} - b^{-1} - c^{-3}}{b^{-1}c^{-2}\omega^{-1} - b^{-1}c^{-3}} = \frac{bc^3 + c\omega^2 - c^3 - b}{c\omega^2 - 1} \\
 &= \omega \left(b(c^2 + c\omega + \omega^2) - c(c + \omega) \right).
 \end{aligned}$$

Finally, we infer

$$\begin{aligned}
 p + \omega q + \omega^2 r &= b^2 + bc + c^2 - b^2c - bc^2 + b^2c + bc\omega^2 \\
 &= c\omega - b^2 - b\omega^2 + bc^2 + bc\omega + b\omega^2 - c^2 - c\omega = 0,
 \end{aligned}$$

and the claim follows from Proposition 4.40. \square

4.4 Summing Up

Complex numbers

Complex numbers are points in the plane \mathbb{R}^2 : one identifies 1 to $(1, 0)$ and denotes by i the number corresponding to $(0, 1)$. Complex addition then coincides with the sum of plane vectors. Any complex number $z = (x, y)$ then is written as $x + iy$ and complex multiplication reduces to standard rules plus $i^2 = -1$.

If $z = x + iy \in \mathbb{C}$, then

$$\begin{aligned}
 \Re(z) &:= x, & \Im(z) &:= y, & \bar{z} &:= x - iy, \\
 \Re(z) &= \frac{z + \bar{z}}{2}, & \Im(z) &= \frac{z - \bar{z}}{2i}, & |z|^2 &:= x^2 + y^2 = z\bar{z}.
 \end{aligned}$$

Polar form

Every complex number $z \neq 0$ appears in *polar form* as $z = |z|(\cos \theta + i \sin \theta)$ where θ , which is defined modulo a multiple of 2π , is called the *argument* of z . We have

- for $z = |z|(\cos \theta + i \sin \theta)$ and $w = |w|(\cos \varphi + i \sin \varphi)$, $zw = |z||w|(\cos(\theta + \varphi) + i \sin(\theta + \varphi))$.
- DE MOIVRE'S FORMULA. $z^n = |z|^n(\cos n\theta + i \sin n\theta)$.
- The argument of a complex number $z \neq 0$ is not uniquely defined. In order to consider the argument as a real function $\arg(z) : \mathbb{C} \setminus \{0\} \rightarrow \mathbb{R}$, we need to choose a *determination*, that is an interval of size 2π in which to read the argument: a common choice is the *principal determination* $\arg : \mathbb{C} \setminus \{0\} \rightarrow [0, 2\pi[$.

The n -th roots

If $w \in \mathbb{C} \setminus \{0\}$, there are n distinct n -th roots of w , i.e., n distinct solutions of the equation $z^n = w$, given by

$$z_j := |w|^{1/n} e^{i \frac{\arg w}{n}} \omega^j, \quad j = 0, 1, \dots, n-1,$$

where $\omega := e^{i \frac{2\pi}{n}}$. The numbers

$$z_j := \omega^j, \quad j = 0, 1, \dots, n-1$$

are the n -th roots of unity, i.e., the solutions of $z^n = 1$.

Complex exponential

Define the *complex exponential* by

$$e^z := e^x(\cos y + i \sin y), \quad \text{for all } z = x + iy \in \mathbb{C}.$$

- We have

$$e^0 = 1, \quad e^{i\pi} = -1, \quad e^{zw} = e^z e^w \quad \forall z, w \in \mathbb{C}.$$

- EULER'S FORMULAS. If $t \in \mathbb{R}$, then

$$e^{it} := \cos t + i \sin t, \quad \cos \omega t = \frac{e^{i\omega t} + e^{-i\omega t}}{2}, \quad \sin \omega t = \frac{e^{i\omega t} - e^{-i\omega t}}{2i}.$$

Complex notation appears as a great simplification of the description of the harmonic functions.

- the uniform circular motion on the unit circle with angular velocity ω passing through 1 at $t = 0$, is described by $t \rightarrow e^{i\omega t}$, $t \in \mathbb{R}$.
- for $\lambda \in \mathbb{C}$ formulas

$$D(e^{i\lambda t}) = i\lambda e^{i\lambda t}, \quad \int_0^t e^{i\lambda s} ds = \frac{e^{i\lambda t} - 1}{i\lambda}, \lambda \neq 0,$$

are handier than $D(e^{at} \cos(bt)) = ae^{at} \cos(bt) - be^{at} \sin(bt)$, and $D(e^{at} \sin(bt)) = ae^{at} \sin(bt) + be^{at} \cos(bt)$, or the corresponding formulas for the primitives.

- PROSTAPHERESIS FORMULAS.

$$\begin{cases} e^{i\alpha} + e^{i\beta} = 2 \cos \frac{\alpha-\beta}{2} e^{i \frac{\alpha+\beta}{2}}, \\ e^{i\alpha} - e^{i\beta} = 2i \sin \frac{\alpha-\beta}{2} e^{i \frac{\alpha+\beta}{2}}. \end{cases}$$

They clearly explain the *beating phenomenon* between two oscillators, even with different amplitudes,

$$c_1 e^{i\omega_1 t} + c_2 e^{i\omega_2 t} = e^{i \frac{\omega_1 + \omega_2}{2} t} \left\{ c_1 e^{i \frac{\omega_1 - \omega_2}{2} t} + c_2 e^{-i \frac{\omega_1 - \omega_2}{2} t} \right\}.$$

4.5 Exercises

4.45 ¶. Write the following complex numbers in polar form:

$$5 - 5i, \quad 1 + i\sqrt{3}, \quad 2 - 5i, \quad 1 - i.$$

4.46 ¶. Determine the points in the complex plane such that

$$\Re \frac{1}{z} = \frac{1}{a}, \quad \Re \frac{z-1}{z+1} = 0, \quad |z-i| + |z+i| < 4, \quad \left| \frac{z-z_1}{z-z_2} \right| = k.$$

Describe algebraically the sets

- (i) of the points that have distance at most 1 from the imaginary axis;
- (ii) of the points in the positive half-plane, with distance at least 2 from the origin.

4.47 ¶. Write in the form $a + ib$ the numbers

$$\frac{2-i}{1+2i}, \quad (1+i)^2, \quad \left(\frac{1+i}{1-i} \right)^2.$$

4.48 ¶. Compute $\left| \frac{1-i}{1+i} \right|$.

4.49 ¶. Verify the following:

$$\cos 3\theta = \cos^3 \theta - 3 \cos \theta \sin^2 \theta,$$

$$\sin 3\theta = 3 \cos^2 \theta \sin \theta - \sin^3 \theta,$$

$$\cos 4\theta = \cos^4 \theta - 6 \cos^2 \theta \sin^2 \theta + \sin^4 \theta = 1 - 8 \cos^2 \theta \sin^2 \theta,$$

$$\sin 4\theta = 4 \cos^3 \theta \sin \theta - 4 \cos \theta \sin^3 \theta.$$

4.50 ¶. Compute

$$(1+i)^4, \quad (3-3i)^4, \quad (-5+5i)^8.$$

4.51 ¶. de Moivre's formula allows us to express $\cos n\theta$ and $\sin n\theta$ by means of $\cos \theta$ and $\sin \theta$. Find those formulas. [*Hint:* Use Newton's binomial.]

4.52 ¶ Fagnano formula. Show that $2i \log \frac{1-i}{1+i} = \pi$.

4.53 ¶. Infer the following equalities from Euler's formulas:

$$\cos^3 x = \frac{1}{4} \cos 3x + \frac{3}{4} \cos x,$$

$$\sin^4 x = \frac{1}{8} \cos 4x = \frac{1}{2} \cos 2x + \frac{3}{8},$$

$$\sin^5 x = \frac{1}{16} \sin 5x - \frac{5}{16} \sin 3x + \frac{5}{8} \sin x.$$

4.54 ¶. Prove that

$$\sin \theta + \sin 2\theta + \sin 3\theta + \cdots + \sin n\theta = \frac{\sin \frac{n+1}{2}\theta}{\sin \frac{\theta}{2}} \sin \frac{n\theta}{2},$$

$$1 + \cos \theta + \cos 2\theta + \cos 3\theta + \cdots + \cos n\theta = \frac{\sin(n + \frac{1}{2})\theta}{2 \sin \frac{\theta}{2}}.$$

[*Hint:* Recall $e^{i\theta} = \cos \theta + i \sin \theta$ and de Moivre's formula.]

4.55 ¶. Compute

$$\sqrt[4]{4}, \sqrt[3]{-1}, \sqrt{i}, \sqrt{2-2i}, \sqrt[4]{1+i\sqrt{3}}, \sqrt[3]{8-8i}, \sqrt[4]{-4}.$$

4.56 ¶¶. Denote by $\epsilon_0, \dots, \epsilon_{n-1}$ the n -th roots of unity. Show that they form a multiplicative group of finite order n , (i.e., only the first $n-1$ roots are distinct. We say that $e \in G_n$ is a *generator* of G_n if the elements $1, e, e^2, \dots, e^{n-1}$ are distinct. Show that e_h is a generator of G_n if and only if h and n are coprime.

4.57 ¶. Show that the sum of the n -th roots of a number is zero.

4.58 ¶. Solve the equations

$$\begin{aligned} z^2 + 3iz + 4 &= 0, & z^2 &= \bar{z}, & z^2 + i\bar{z} &= 1, \\ z^2 + 2z + i &= 0, & (z+i)^3 &= (\sqrt{3}+i)^3, \\ z|z| - 2z - 1 &= 0, & z^4 &= \bar{z}^3, & z^3 &= iz\bar{z}, \\ |z|^2 z^2 &= i, & \Re z^4 &= |z|^4, & z|z| - 2\Re z &= 0, \\ z^2 + z\bar{z} &= 1 + 2i, & \Re z &= \frac{1}{2}|z|^2. \end{aligned}$$

4.59 ¶. Show that z and cz are orthogonal in $\mathbb{R}^2 \sim \mathbb{C}$ if and only if c is purely imaginary.

4.60 ¶. Verify that

$$\begin{aligned} \log(-5) &= \log 5 + i\pi, \\ \log(-\sqrt{3} + i) &= \log 2 + i\frac{5}{6}\pi, \\ \log(7 + 7i) &= \log 7 + \frac{1}{2}\log 2 + i\frac{\pi}{4}. \end{aligned}$$

4.61 ¶. Interpret in complex notation the *two-squares theorem* of Diophantus of Alexandria (200–284)

$$(u^2 + v^2)(x^2 + y^2) = (ux - vy)^2 + (uy + vx)^2.$$

4.62 ¶. Let z_1, z_2, z_3, z_4 be four points on a circle centered at the origin. Show that the following claims are equivalent:

- (i) z_1, z_2, z_3, z_4 are the vertices of a rectangle,
- (ii) $z_1 + z_2 + z_3 + z_4 = 0$,
- (iii) z_1, z_2, z_3, z_4 are the roots of an equation of the type $(z^2 - a^2)(z^2 - b^2) = 0$ with $|a| = |b| \neq 0$.

4.63 ¶¶. \circ $f : \mathbb{C} \rightarrow \mathbb{C}$ is said to be an *isometry* or is *distance preserving* if $|f(w) - f(z)| = |z - w| \forall z, w \in \mathbb{C}$. Show that f is distance preserving if and only if $f(z) := f(0) + bz$ or $f(z) = f(0) + b\bar{z}$ with $|b| = 1$.

\circ $f : \mathbb{C} \rightarrow \mathbb{C}$ is said to be *\mathbb{R} -linear* if $f(z) = \Re z f(1) + \Im z f(i)$. Show that f is \mathbb{R} -linear if and only if $f(z) = az + b\bar{z}$ with $a, b \in \mathbb{C}$.

\circ An \mathbb{R} -linear map $f : \mathbb{C} \rightarrow \mathbb{C}$ is said to be *orthogonal* if $\Re(f(z)\overline{f(w)}) = \Re(z\bar{w})$ for all $z, w \in \mathbb{C}$. Show that f is orthogonal if and only if $f(z) = az$ or $f(z) = a\bar{z}$ with $a \in \mathbb{C}$ and $|a| = 1$.

[Hint: For the first claim, consider $g(z) = (f(z) - f(0))/(f(1) - f(0))$, and show that $|g(z)|^2 = |z|^2$ and $|g(z) - 1|^2 = |z - 1|^2$.]

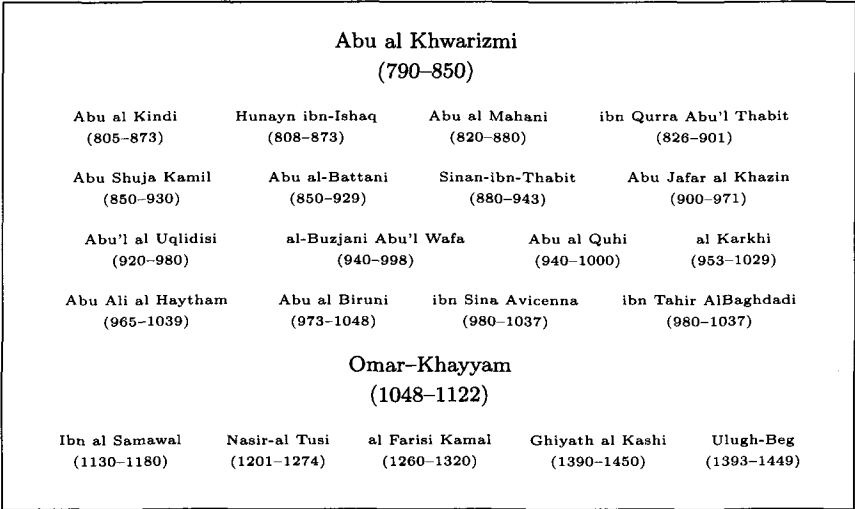


Figure 4.12. The arab renaissance.

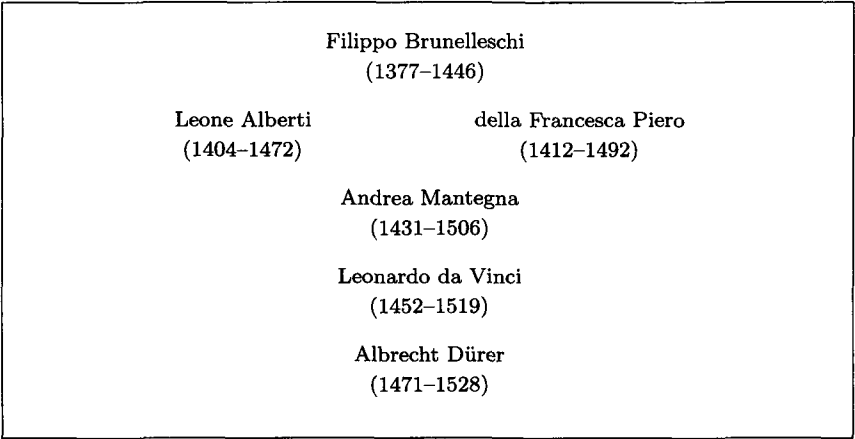


Figure 4.13. Mathematics and art in the Renaissance period.

5. Polynomials, Rational Functions and Trigonometric Polynomials

In this chapter we want to illustrate the relevance of complex numbers in some elementary situations. After a brief discussion of the algebra of polynomials in Section 5.1, we prove the *fundamental theorem of algebra* and discuss solutions by radicals of algebraic equations in Section 5.2. In Section 5.3 we present *Hermite's decomposition formulas* for rational functions, which are useful for the integration of rational functions, see Chapter 4 of [GM1]. Finally, in Section 5.4, we discuss some basic facts about *trigonometric polynomials* and, more generally, sums of sinusoidal signals. In particular we shall see that the *spectrum* of a signal completely identifies the signal itself, we shall prove the *energy identity* and present a *sampling formula*.

5.1 Polynomials

Let \mathbb{K} be a field as, for instance, \mathbb{C} , \mathbb{R} , \mathbb{Q} , or a finite field as, for example, the residue class \mathbb{Z}_p , p prime. A *polynomial* with coefficients in \mathbb{K} in the *indeterminate* x is an expression of the form

$$P(x) := a_0 + a_1x + a_2x^2 + \cdots + a_px^p = \sum_{j=0}^p a_jx^j, \quad ^1$$

where $a_j \in \mathbb{K}$ for all $j = 0, \dots, p$. The class of all polynomials with coefficients in \mathbb{K} in the indeterminate x will be denoted by $\mathbb{K}[x]$.

Presently a polynomial $P(x) \in \mathbb{K}[x]$ is not a function defined in some domain, but, instead, a formal expression defined essentially by the list of its coefficients. In fact we say that *two polynomials* $P(x) = \sum_{j=0}^n a_jx^j$ and $Q(x) = \sum_{j=0}^m b_jx^j$ are equal if $a_j = b_j \forall j = 0, \dots, \min(n, m)$ and $a_j = 0 \forall j = \min(n, m) + 1, \dots, n$ and $b_j = 0 \forall j = \min(n, m) + 1, \dots, m$.

We can therefore extend, if this is convenient, the list of the coefficients of a polynomial by adding zeros as coefficients of higher order terms without changing the polynomial itself.

¹ $a_0 + \sum_{j=1}^p a_jx^j$ is the actual meaning of the sum!

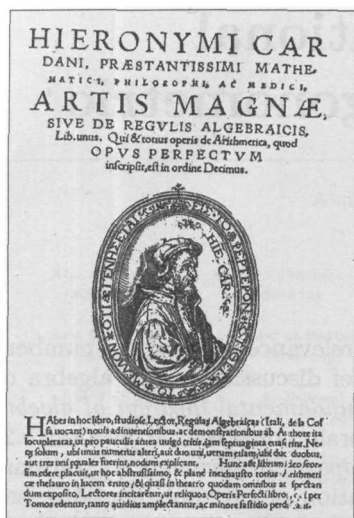


Figure 5.1. The frontispieces of *Ars Magna* by Girolamo Cardano (1501–1576) and of *Algebra* by Rafael Bombelli (1526–1573).

If $P(x) = \sum_{j=0}^n a_j x^j \in \mathbb{K}[x]$, the largest integer j for which $a_j \neq 0$ is called the *degree of P* and is denoted by $\deg P$. Nonzero constant polynomials have degree 0, and the zero-polynomial, that is the polynomial with $a_j = 0 \forall j$, is given degree $-\infty$.

Polynomials in $\mathbb{K}[x]$ can be added and multiplied. If $P(x) = \sum_{j=0}^p a_j x^j$ and $Q(x) = \sum_{j=0}^q b_j x^j \in \mathbb{K}[x]$, and assuming for instance $p \geq q$, we define the sum of P and Q by

$$P(x) + Q(x) := \sum_{j=0}^p (a_j + b_j) x^j$$

where we have set $b_j = 0$ for $j = q + 1, \dots, p$. Of course $\deg(P + Q) \leq \max(\deg P, \deg Q)$. The product of P and Q is then defined by

$$P(x)Q(x) = \left(\sum_{i=0}^p a_i x^i \right) \left(\sum_{j=0}^q b_j x^j \right) = \sum_{i=0}^p \sum_{j=0}^q a_i b_j x^i x^j := \sum_{k=0}^{p+q} c_k x^k$$

where

$$c_k := \sum_{\substack{i,j \\ i+j=k}} a_i b_j$$

or, explicitly,

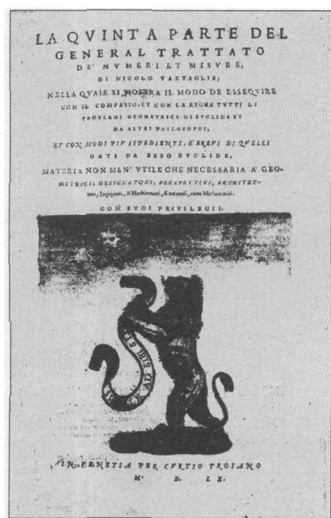


Figure 5.2. Frontispieces of the first and fifth parts of the *General Trattato sui numeri* of Niccolò Fontana (1500–1557), called Tartaglia.

$$\begin{cases} c_0 = a_0 b_0, \\ c_1 = a_1 b_0 + a_0 b_1, \\ c_2 = a_2 b_0 + a_1 b_1 + a_0 b_2, \\ \dots \\ c_n = a_n b_0 + a_{n-1} b_1 + a_{n-2} b_2 + \dots + a_1 b_{n-1} + a_0 b_n, \\ \dots \end{cases}$$

Notice that we have extended the list of the coefficients of the polynomials by setting $a_j := 0$ for $j = p+1, \dots, p+q$, and $b_j := 0$ for $j = q+1, \dots, p+q$. It is easy to see that $\deg(PQ) = \deg P + \deg Q$.

5.1 ¶. Show that the product of two polynomials is zero if and only if one of the two polynomials is zero. This is expressed by saying that $\mathbb{K}[x]$ is an *integral domain*.

5.1.1 The Division Algorithm

Given two polynomials $A(x) = \sum_{j=0}^n a_j x^j$ and $B(x) = \sum_{j=0}^m b_j x^j \in \mathbb{K}[x]$ with $\deg B = m \leq n$, we observe that

$$A_1(x) := A(x) - \frac{a_n}{b_m} x^{n-m} B(x) \in \mathbb{K}[x]$$

has degree less than n . Proceeding inductively, it is not difficult to show that

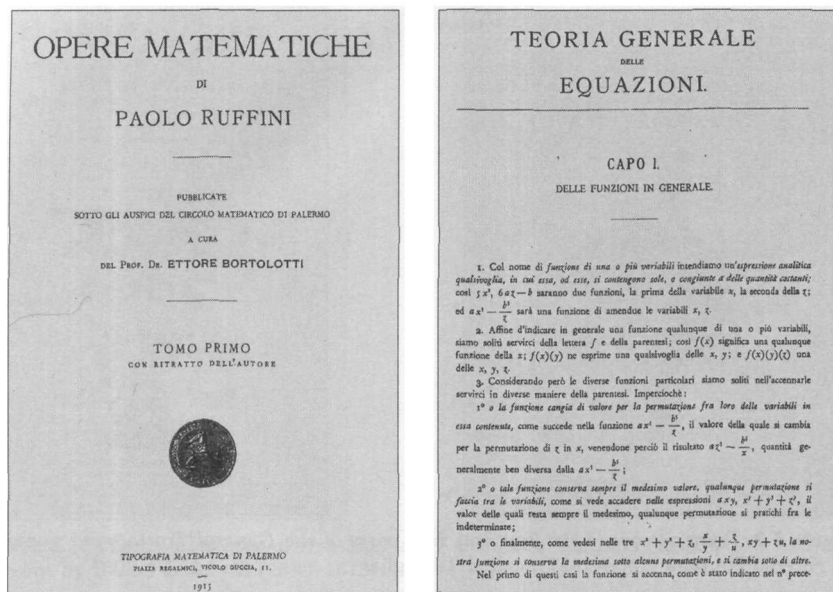


Figure 5.3. The frontispiece of the *Opere Matematiche* by Paolo Ruffini (1765–1822) and the first page of the first chapter of the *Teoria generale delle equazioni in cui si dimostra impossibile la soluzione algebrica delle equazioni di grado superiore al quarto* by Paolo Ruffini (1765–1822).

5.2 Theorem (Division algorithm). For given $A, B \in \mathbb{K}[x]$ with $B \neq 0$, there are uniquely defined polynomials $Q, R \in \mathbb{K}[x]$ such that

$$A = BQ + R \quad \text{and} \quad \deg R < \deg B.$$

The polynomial Q in Theorem 5.2 is called the *integral quotient* of A by B , and R is called the *remainder* of A divided by B .

a. Euclid's algorithm and Bezout identity

Let $A, B \in \mathbb{K}[x]$ be two nonzero polynomials. We say that B *divides* A or that B is a *divisor* of A if $A = BQ$ for some $Q \in \mathbb{K}[x]$. Notice that, if B is a divisor of A , then λB for $\lambda \in \mathbb{K}$, $\lambda \neq 0$, is a divisor of A , too. Notice that this contrasts with the notion of integral divisor of an integral number. We shall say that A is *irreducible in \mathbb{K}* if it has no divisors. In this case if $A = BQ$, then either B or Q reduces to a constant polynomial.

We now look for the common divisors of A and B . Clearly polynomials of degree zero are common divisors of A and B ; also, every common divisor to A and B divides $PA + QB$ for all polynomials P, Q .

We say that a subset $\mathcal{I} \subset \mathbb{K}[x]$ is an *ideal* of $\mathbb{K}[x]$ if \mathcal{I} is a subgroup with respect to the addition, and it is closed with respect to the multiplication with any element, that is, $PQ \in \mathcal{I} \forall P \in \mathbb{K}[x], \forall Q \in \mathcal{I}$.

5.3 Proposition. *Every nonzero ideal \mathcal{I} of polynomials with one indeterminate is a principal ideal, that is, it contains only multiples (in $\mathbb{K}[x]$) of a polynomial that is unique up to a multiplication by a constant.*

Proof. Let B be a polynomial in \mathcal{I} with minimal degree and let A be any other element in \mathcal{I} . By dividing A by B we have $A = BQ + R$, thus $R = A - BQ$ belongs to \mathcal{I} . Since $\deg R < \deg B$, we conclude that $R = 0$, i.e., $A = BQ$. Suppose now that $B' \in \mathcal{I}$ and $\deg B' = \deg B$; $B' = BQ$ yields $\deg Q = 0$, i.e., $B' = \lambda B$, $\lambda \in \mathbb{C}$. \square

Since

$$\mathcal{I} := \left\{ PA + QB \mid P, Q \in \mathbb{K}[x] \right\}$$

is an ideal of $\mathbb{K}[x]$, we conclude that there exists a polynomial D , uniquely defined modulo a multiplicative constant, such that every polynomial $PA + QB$ in \mathcal{I} is a multiple of D . In particular

- $D = AP + BQ$ for some P, Q ,
- $A = \bar{A}D$ and $B = \bar{B}D$, since $1 \cdot A + 0 \cdot B, 0 \cdot A + 1 \cdot B \in \mathcal{I}$.

We therefore conclude that every common divisor of A and B divides D and that every divisor of D divides both A and B . That is, D is the (up to a multiplicative constant) *greatest common divisor* of A and B . With some abuse of notation, it is denoted by $\text{g.c.d.}(A, B)$.

As for integers, *Euclid's algorithm* and *Euclid's generalized algorithm* yield a way to compute the greatest common divisor of A and B together with polynomials U, V such that $AU + BV = \text{g.c.d.}(A, B)$. Assume $\deg A \geq \deg B$ and define the three sequences $\{R_k\}$, $\{U_k\}$, $\{V_k\}$ by

$$\begin{cases} R_0 := A, R_1 := B, \\ R_{k+1} = R_{k-1} - Q_k R_k, \\ U_0 := 1, U_1 := 0, \\ U_{k+1} = U_{k-1} - Q_k U_k, \\ V_0 := 0, V_1 := 1, \\ V_{k+1} = U_{k-1} - Q_k V_k, \end{cases}$$

until $R_{n+1} \neq 0$.

5.4 Theorem (Euclid). *We have $R_n := \text{g.c.d.}(A, B)$, and*

$$\text{g.c.d.}(A, B) = R_n = AU_n + BV_n.$$

Proof. In fact, noticing that C divides A and B if and only if C divides B and the remainder $R := A - BQ$, by induction one proves

$$\begin{aligned}\text{g.c.d.}(A, B) &= \text{g.c.d.}(R_0, R_1) = \text{g.c.d.}(R_1, R_2) \\ &= \cdots = \text{g.c.d.}(R_{n-1}, R_n) = R_n.\end{aligned}$$

(ii) It suffices to check by induction that $R_k = AU_k + BV_k \forall k = 0, \dots, n$. \square

b. Factorization

Two polynomials that have a constant polynomial as greatest common divisor are said to be *coprime*, and a polynomial is said to be *prime* or *irreducible* over \mathbb{K} if it has no divisors except for the nonzero constants.

From Euclid's algorithm, it is easy to see that $\text{g.c.d.}(AC, BC) = \text{g.c.d.}(A, B)C$. This implies the following.

5.5 Theorem (Euclid). *If A divides BC and A and B are coprime, then A divides C .*

This, in turn, allows us to prove as for integers the following.

5.6 Theorem (Unique factorization). *Every polynomial in $\mathbb{K}[x]$ can be uniquely written as a product of irreducible factors.*

Thus irreducible factors play in $\mathbb{K}[x]$ the same role as prime numbers in arithmetic.

5.7 Remark. We notice that the notion of irreducible polynomial depends on the field \mathbb{K} of coefficients. For instance, $x^2 - 4$ is not prime in $\mathbb{Q}[x]$, $x^2 - 2$ is prime in $\mathbb{Q}[x]$ but not in $\mathbb{R}[x]$, nor in $\mathbb{C}[x]$, and $x^2 + 1$ is prime in $\mathbb{R}[x]$ but not in $\mathbb{C}[x]$.

c. The factor theorem

A polynomial $P(x) = \sum_{j=0}^n a_j x^j \in \mathbb{K}[x]$ may be also regarded as a function $P : \mathbb{K} \rightarrow \mathbb{K}$ which maps $z \in \mathbb{K} \rightarrow P(z) := \sum_{j=0}^n a_j z^j$, which we call the *polynomial function* of P . In general, two different polynomials may have the same polynomial function, for instance $x + 1$ and $x^3 + 1$ on \mathbb{Z}_2 , but, as we shall see in a moment, two polynomials are identical if and only if they have the same polynomial function, provided the field \mathbb{K} is infinite.

We say that $\alpha \in \mathbb{K}$ is a *zero* of a polynomial $P \in \mathbb{K}[x]$ if $P(\alpha) = 0$. From the division algorithm theorem we infer at once

5.8 Theorem (Ruffini). *Let $P(x) = \sum_{j=0}^p a_j x^j \in \mathbb{K}[x]$ be a polynomial of degree p and let $\alpha \in \mathbb{K}$. Then $P(x) = (x - \alpha)Q_\alpha(x) + P(\alpha) \forall x \in \mathbb{K}$ where $Q_\alpha(x) = \sum_{j=0}^{p-1} b_j x^j$ with*

$$\begin{cases} b_{p-1} = a_p, \\ b_{j-1} = a_j + \alpha b_j \text{ (in } \mathbb{K}, \quad \forall j = p-1, p-2, \dots, 1. \end{cases}$$

Proof. In fact,

$$\begin{aligned}(x - \alpha)Q_\alpha(x) &= \sum_{j=1}^{p-1} b_j x^{j+1} - \alpha \sum_{j=0}^{p-1} b_j x^j = b_{p-1} x^p + \sum_{j=1}^{p-1} (b_{j-1} - \alpha b_j) x^j - \alpha b_0 \\ &= a_p x^p + \sum_{j=1}^{p-1} a_j x^j - \alpha b_0 = P(x) - (a_0 - \alpha b_0).\end{aligned}$$

□

Theorem 5.8 yields the following.

5.9 Theorem (Factor theorem). *Let $P \in \mathbb{K}[x]$ and $\alpha \in \mathbb{K}$. Then $x - \alpha$ divides $P(x)$ if and only if α is a root of P , $P(\alpha) = 0$.*

If we know k distinct roots x_1, x_2, \dots, x_k of P , we can write inductively $P(x) = (x - x_1)Q_1(x)$, $P(x) = (x - x_1)(x - x_2)Q_2(x)$ as $Q_1(x_2) = 0, \dots$, and finally

$$P(x) = (x - x_1)(x - x_2) \cdots (x - x_k)Q_k(x), \quad (5.1)$$

where $\deg Q_k = \deg P - k$. In particular we cannot exceed $\deg P$, that is, *every polynomial of degree n has at most n roots.*

5.10 Theorem (Principle of identity of polynomials). *Two polynomials P and $Q \in \mathbb{K}[x]$ of degree at most n are equal in $\mathbb{K}[x]$ if and only if their polynomial functions take equal values in at least $n + 1$ distinct points of \mathbb{K} . In particular, if P has degree n and its polynomial function vanishes in at least $n + 1$ distinct points, then $P = 0$ in $\mathbb{K}[x]$.*

5.11 $P(x)$ in the indeterminate $x - \alpha$. By using the factor theorem, it is easy to rewrite $P \in \mathbb{K}[x]$ as a polynomial in the indeterminate $x - \alpha$. We have the following.

Proposition. *Let P be a polynomial of degree p and let $\alpha \in \mathbb{K}$. Let Q_p, Q_{p-1}, \dots, Q_1 be the polynomials of degrees respectively $p, p-1, \dots, 2, 1$ obtained iteratively by Ruffini's rule, i.e.,*

$$\begin{cases} Q_n := P, \\ Q_j(z) - Q_j(\alpha) = (z - \alpha)Q_{j-1}(z) \quad \forall j = n-1, \dots, 2, 1. \end{cases}$$

Then $P(x) = P(\alpha) + \sum_{j=1}^p Q_{p-j}(\alpha)(x - \alpha)^j$.

Proof. In fact,

$$\begin{aligned}Q_0(x) &= Q_0(\alpha), \\ Q_1(x) &= Q_1(\alpha) + (x - \alpha)Q_0(x), \\ Q_2(x) &= Q_2(\alpha) + (x - \alpha)Q_1(x) = Q_2(\alpha) + (x - \alpha)Q_1(\alpha) + (x - \alpha)^2 Q_0(\alpha), \\ &\dots\end{aligned}$$

$$P(x) = Q_p(x) = \sum_{j=0}^p Q_{p-j}(\alpha)(x - \alpha)^j.$$

□

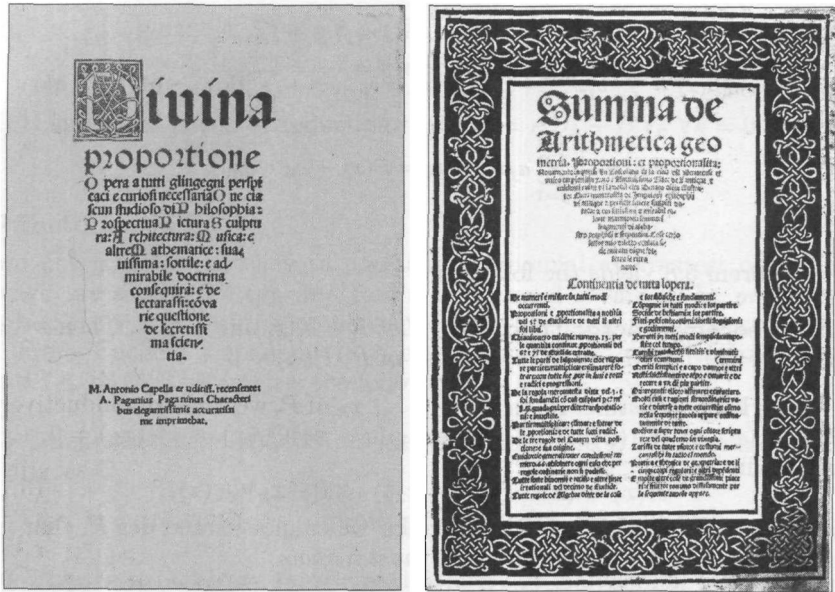


Figure 5.4. The Italian Renaissance is probably the turning point for the development of modern western culture. The knowledge of artists, technicians and merchants merges with school education. Luca Pacioli (1445–1517) writes for “curious and ingenious engineers and for any scholar of philosophy, perspective, painting, sculpture, architecture, music and other mathematics.”

5.12 Complex derivative. Let $P(z) = \sum_{j=0}^n a_j(z - \alpha)^j \in \mathbb{C}[z]$. The *complex derivative* of P is defined as the polynomial of degree $n - 1$,

$$DP(z) = P'(z) := \sum_{j=1}^n j a_j (z - \alpha)^{j-1}.$$

Of course $D^2P(z) := D(DP)(z) = \sum_{j=2}^n j(j-1)a_j z^{j-2}$, and

$$D^k P(z) = \begin{cases} \sum_{j=k}^n j(j-1) \cdots (j-k+1) a_j (z - \alpha)^{j-k} & \text{if } k \leq n, \\ 0 & \text{if } k > n. \end{cases}$$

In particular for $0 \leq k \leq n$, $D^k A(\alpha) = k! a_k$. Consequently we infer, compare Taylor’s formula in [GM1],

$$P(z) = \sum_{j=0}^n \frac{D^j P(\alpha)}{j!} (z - \alpha)^j. \quad (5.2)$$



Figure 5.5. The famous architect Leon Battista Alberti (1404–1472) was the theorist of mathematical perspective. His ideas were presented in *De Pictura*, 1511, while in *Ludi Mathematici* he discussed applications of mathematics to various practical problems. The frontispiece of his *De re aedificatoria*, organic summa of the architecture of his time.

5.1.2 The fundamental theorem of algebra

Finding the irreducible polynomials in $\mathbb{K}[x]$ is obviously a key point of the algebra of polynomials. We shall restrict ourselves to factorization in the fields \mathbb{R} and \mathbb{C} .

a. Factorization in \mathbb{C}

Let $P(z) = \sum_{k=0}^n a_k z^k$ be a polynomial in $\mathbb{C}[z]$ of degree n , i.e., $a_n \neq 0$. Trivially every root of $P(z)$, that is a point $\xi \in \mathbb{C}$ such that $P(\xi) = 0$, is a minimum point for the real-valued function $z \rightarrow |P(z)|$. An interesting fact discovered by Jean d'Alembert (1717–1783) and Jean Argand (1768–1822) is that the converse holds true.

5.13 Proposition. *Let ξ be a local minimizer for $|P(z)|$, i.e., there is a disk $B(\xi, \rho)$ of radius ρ and center ξ such that $|P(\xi)| \leq |P(z)| \forall z \in B(\xi, \rho)$; then $|P(\xi)| = 0$.*

This is a consequence of the following.

5.14 Lemma (d'Alembert). *Let $z_0 \in \mathbb{C}$ be such that $P(z_0) \neq 0$; then for any $\epsilon > 0$ we can find $h \in \mathbb{C}$ with $|h| < \epsilon$ such that $|P(z_0 + h)| < |P(z_0)|$.*

In order to prove d'Alembert's lemma we make the following remarks.

- Let $P(z) = \sum_{j=0}^n a_j z^j$ and let $\alpha \in \mathbb{C}$. As we have seen, in 5.11, we can write $P(z)$ as

$$P(z) - P(\alpha) = \sum_{j=1}^p A_j (z - \alpha)^j$$

where the coefficients A_j depend on the coefficients of P and α but not on z . Denote by m the smallest integer such that $A_j \neq 0$. Clearly $1 \leq m \leq p$ and

$$P(z) - P(\alpha) = \sum_{j=m}^p A_j (z - \alpha)^j. \quad (5.3)$$

- From (5.3) and the triangle inequality, we infer

$$|P(z) - P(\alpha)| = \left| \sum_{j=m}^p A_j (z - \alpha)^j \right| \leq \sum_{j=m}^p |A_j| |z - \alpha|^j,$$

therefore for all z with $|z - \alpha| < 1$ we have

$$|P(z) - P(\alpha)| \leq k |z - \alpha|^m \quad k := \sum_{j=m}^p |A_j|. \quad (5.4)$$

Proof of Lemma 5.14. According to the above, for $h \in \mathbb{C}$ we write

$$P(z_0 + h) = P(z_0) + \sum_{j=1}^n A_j h^j, \quad (5.5)$$

and, if m is the smallest integer with $A_m \neq 0$ and

$$Q(h) := \sum_{j=m+1}^n A_j h^j,$$

we rewrite (5.5) as

$$P(z_0 + h) = P(z_0) + A_m h^m + Q(h).$$

We now choose h_0 in such a way that $A_m h_0^m$ is in the opposite direction of $P(z_0)$,

$$A_m h_0^m = -P(z_0), \quad (5.6)$$

i.e., we choose h_0 as one of the m -th roots of $-P(z_0)/A_m$, which is possible, since we are working in \mathbb{C} . Then we set $h = \rho h_0$, ρ small and precisely $\rho|h_0| = |h| < 1$. From (5.6) we then infer

$$|Q(h)| \leq k|h|^{m+1} = \frac{k|h_0|}{|A_m|} |A_m| |h_0|^m \rho^{m+1} = \frac{k|h_0|}{|A_m|} \rho^{m+1} |P(z_0)|$$

hence



Figure 5.6. Jean d'Alembert (1717–1783) and Carl Friedrich Gauss (1777–1855).

$$|Q(h)| < \frac{1}{2}\rho^m|P(z_0)|,$$

if, moreover, $\rho < |A_m|/(2k|h_0|)$. Finally, from (5.5) and the triangle inequality we conclude

$$\begin{aligned} |P(z_0 + h)| &\leq (1 - \rho^m)|P(z_0)| + |Q(h)| \\ &\leq (1 - \rho^m + \frac{1}{2}\rho^m)|P(z_0)| < |P(z_0)|. \end{aligned}$$

□

Besides Proposition 5.13 we also have

5.15 Lemma (Coercivity). *Let $P(z) = \sum_{k=0}^n a_k z^k$ be a polynomial of degree $n \geq 1$. Then*

$$\lim_{|z| \rightarrow \infty} |P(z)| = +\infty.$$

Proof. Factoring out the term of highest degree, we have

$$P(z) = a_n z^n (1 + Q(1/z)), \quad Q(w) := \sum_{j=1}^n (a_{n-j}/a_n) w^j.$$

If $k := \sum_{j=1}^n |a_{n-j}/a_n|$, $|z| > 1$ and $|z| > 2k$, applying (5.4) to Q we get

$$|Q(1/z)| = |Q(1/z) - Q(0)| \leq \frac{k}{|z|} \leq \frac{1}{2}.$$

Consequently, using the triangle inequality $|1 + q| \geq |1 - |q||$,

$$\begin{aligned} |P(z)| &= |a_n||z|^n |1 + Q(1/z)| \geq |a_n||z|^n |1 - |Q(1/z)|| \\ &= |a_n||z|^n (1 - |Q(1/z)|) \geq \frac{1}{2}|a_n||z|^n; \end{aligned}$$

this yields the result at once.

□

every polynomial $P \in \mathbb{C}[z]$ of degree n has exactly n roots, when counted with their multiplicities, i.e.,

$$P(z) = a_n(z - z_1)^{n_1}(z - z_2)^{n_2} \cdots (z - z_k)^{n_k}, \quad n_1 + n_2 + \cdots + n_k = n.$$

The roots of a polynomial and of its derivative are related. It is easy to prove the following:

- A simple root of a polynomial P is not a root of its derivative P' ; consequently P and P' are coprime if and only if all the roots of P are simple.
- A root of multiplicity k of a polynomial is a root of its derivative of multiplicity $k - 1$.
- Let P be a polynomial. Then $P_0 := P/\text{g.c.d.}(P, P')$ has the same set of zeros of P but all its roots are simple.

The following claims are easy consequences of Rolle's theorem:

- If all roots of a real polynomial are real, then all roots of its derivative are also real.
- If all roots of a real polynomial P are real and of those p are positive, then P' has p or $p - 1$ positive roots.

c. Factorization in \mathbb{R}

If $P(z) = \sum_{k=0}^n a_k z^k \in \mathbb{C}[z]$, the conjugate polynomial is defined by $\overline{P(z)} := \sum_{k=0}^n \overline{a_k} z^k$. Of course

$$\overline{P(z)} = \sum_{k=0}^n \overline{a_k} \overline{z}^k = \overline{P(\overline{z})}.$$

It follows: α is a root of P with multiplicity h if and only if $\overline{\alpha}$ is a root for \overline{P} of multiplicity h . Since $P = \overline{P}$ for polynomials with real coefficients, we deduce

5.19 Proposition. *Every real polynomial has n complex roots when counted with their multiplicities; an even number of them are nonreal and come in couples of conjugate complex numbers.*

As a corollary, on account of the fundamental theorem of algebra, we have proved again that *every real polynomial with odd degree has at least one real root.*

Let P be a polynomial of degree n with real coefficients and let $\alpha_1, \alpha_2, \dots, \alpha_p$ be its real roots with respective multiplicities k_1, k_2, \dots, k_p . Moreover, let $\beta_1, \beta_2, \dots, \beta_q$ be its complex roots with positive imaginary parts and multiplicities h_1, h_2, \dots, h_q . Since also $\overline{\beta_1}, \overline{\beta_2}, \dots, \overline{\beta_q}$ are roots of P with multiplicities h_1, h_2, \dots, h_q , we find, on account of the fundamental theorem of algebra, that $k_1 + \cdots + k_p + 2h_1 + \cdots + 2h_q = n$,

$$P(z) = a_n \prod_{j=1}^p (z - \alpha_j)^{k_j} \prod_{j=1}^q (z - \beta_j)^{h_j} (z - \bar{\beta}_j)^{h_j},$$

and, writing $\beta_j =: b_j + ic_j$, so that

$$(z - \alpha_j)(z - \bar{\alpha}_j) = (x - b_j)^2 + c_j^2,$$

we conclude

$$P(x) = a_n \prod_{j=1}^p (x - \alpha_j)^{k_j} \prod_{j=1}^q ((x - b_j)^2 + c_j^2)^{h_j} \quad \forall x \in \mathbb{R}.$$

We therefore have the following.

5.20 Proposition. *Every real polynomial can be factorized as a product of first and second order irreducible polynomials in \mathbb{R} .*

5.2 Solutions of Polynomial Equations

The Italian Renaissance marks a tremendous renewal of interest in nature, and also in mathematics. Artists studied and employed mathematics intensively, among them let us mention Filippo Brunelleschi (1377–1446), Paolo Uccello (1397–1475), Masaccio (1401–1428), Leon Battista Alberti (1404–1472), and Piero della Francesca (1410–1492), who set forth the mathematical principle of perspective. The development of banking and commercial activities called for an improved arithmetic. The *Summa* by Luca Pacioli (1445–1517) and the *General trattato dei numeri e misure* by Niccolò Fontana (1500–1557), called Tartaglia, contained many problems on what one could call numerable mathematics.

With respect to the topic we are discussing, the new flourishing of mathematical studies led to the discovery of formulas for solving algebraic equations of degree 3 and 4 and the consequent introduction of the imaginary unity. These developments are connected to the names of Scipione del Ferro (1465–1526), Niccolò Fontana (1500–1557), called Tartaglia, Girolamo Cardano (1501–1576) and Rafael Bombelli (1526–1573) and are presented in Cardano's *Ars Magna* and Bombelli's *Algebra*.

5.2.1 Solutions by radicals

Equations of second degree

$$az^2 + bz + c = 0, \quad a, b, c \in \mathbb{C}, \quad a \neq 0, \quad (5.7)$$

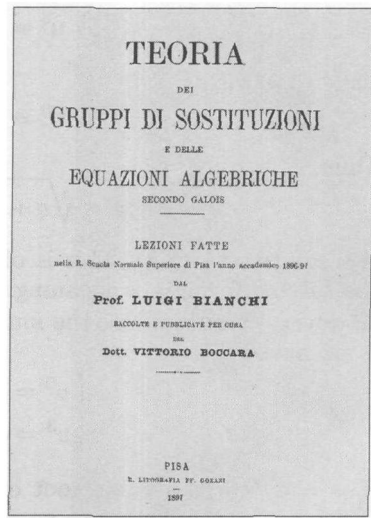
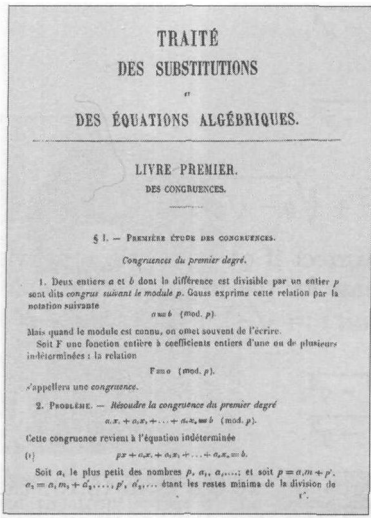


Figure 5.8. The first page of *Traité des substitutions et des équations algébriques* by Camille Jordan (1838–1922) and of *Teoria dei gruppi di sostituzioni* by Luigi Bianchi (1856–1928).

can be easily solved in \mathbb{C} as everybody knows. In fact, completing the square we get

$$a\left(z^2 + 2\frac{b}{2a}z + \frac{b^2}{4a^2}\right) + c - \frac{b^2}{4a} = 0,$$

hence

$$\left(2a\left(z + \frac{b}{2a}\right)\right)^2 + 4ac - b^2 = 0.$$

Thus solutions are given by

$$z_{1,2} = \frac{-b \pm w}{2a}$$

where w and $-w$ are the roots of $w^2 = b^2 - 4ac$.

5.21 Third degree equations. Complex numbers were introduced as an intermediate step to solve the third degree equation

$$x^3 - 3px - 2q = 0, \quad p, q \in \mathbb{R}, \quad p \neq 0, \quad (5.8)$$

that, as we know, has at least a real solution. Set $x = u + v$, $p = uv$, so that (5.8) becomes

$$u^3 + v^3 - 2q = 0 \quad (5.9)$$

hence

$$u^6 - 2qu^3 + p^3 = 0,$$

since $v = p/u$. The last equation is solved by

$$u^3 = q + \sqrt{q^2 - p^3},$$

while (5.9) yields

$$v^3 = q - \sqrt{q^2 - p^3}.$$

Thus

$$x = u + v = \sqrt[3]{q + \sqrt{q^2 - p^3}} + \sqrt[3]{q - \sqrt{q^2 - p^3}}$$

is a solution of (5.8). This is of course correct if $q^2 - p^3 \geq 0$, otherwise the solving formula is meaningless (at least for the Renaissance people). However, if we introduce the *imaginary unit* $i := \sqrt{-1}$ in the case $q^2 - p^3 < 0$, we have

$$\begin{cases} u^3 = q + i\sqrt{p^3 - q^2}, \\ v^3 = q - i\sqrt{p^3 - q^2}. \end{cases} \quad (5.10)$$

If $u = a + ib$ is a cubic root of $q + i\sqrt{p^3 - q^2}$, we see from (5.10) that $v := a - ib$ is a cubic root of $q - i\sqrt{p^3 - q^2}$, therefore the imaginary parts cancel if we sum $u + v$, finding a real root.

5.22 Example. If we consider the equation

$$x^3 - 15x - 4 = 0, \quad p = 5, \quad q = 2, \quad (5.11)$$

we find $x = \sqrt[3]{2 + 11i} + \sqrt[3]{2 - 11i}$. If we try to express $2 + 11i$ as the cube of a complex number, we find $(2 + i)^3 = 8 + 12i + 6i^2 + i^3 = 2 + 11i$, while $(2 - i)^3 = 2 - 11i$, hence $x = 2 + i + 2 - i = 4$ is a solution of (5.11).

An adjustment of the method just presented allows us to solve third degree equations. We want to solve in \mathbb{C} ,

$$P(z) = a_0 z^3 + a_1 z^2 + a_2 z + a_3 = 0, \quad a_0, a_1, a_2, a_3 \in \mathbb{C}, \quad a_0 \neq 0. \quad (5.12)$$

We see that $P''(\alpha) = 0$, if $\alpha := -a_1/(3a_0)$, hence

$$P(z) = a_0(z - \alpha)^3 + P'(\alpha)(z - \alpha) + P(\alpha),$$

and z solves (5.12) if and only if $y := z - \alpha$ solves $a_0 y^3 + P'(\alpha)y + P(\alpha) = 0$. Therefore it suffices to solve equations of the form

$$z^3 + pz + q = 0, \quad p, q \in \mathbb{C}. \quad (5.13)$$

The idea is to look for solutions of the form $z = u + v$. Inserting $z = u + v$ in (5.13), we see that z is a solution if u and v satisfy

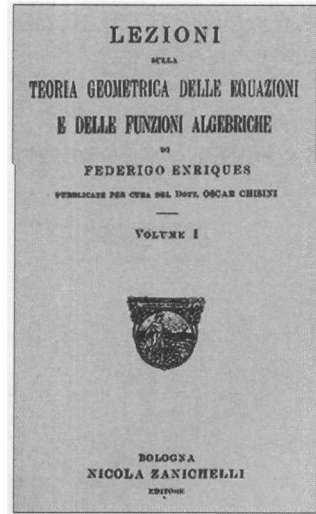
$$\begin{cases} uv = -p/3, \\ u^3 + v^3 = -q, \end{cases}$$

or

$$u^3 v^3 = -p^3/27, \quad u^3 + v^3 = -q.$$



Figure 5.9. Evariste Galois (1811–1832) and the frontispiece of the *Teoria geometrica delle equazioni* by Federigo Enriques (1871–1946).



This happens if u^3 and v^3 are the two solutions r_1, r_2 of the second degree equation

$$r^2 + qr - p^3/27 = 0.$$

The numbers u and v are then to be chosen among the cubic roots of r_1 and r_2 . Set

$$\begin{cases} u_0 := |r_1|^{1/3} e^{i \frac{\arg r_1}{3}}, \\ u_0 := |r_2|^{1/3} e^{i \frac{\arg r_2}{3}}, \\ \omega := -1/2 + i\sqrt{3}/2. \end{cases}$$

Then the solutions of (5.13) are among the numbers

$$z = u_0 \omega^i + v_0 \omega^j, \quad i, j = 0, 1, 2.$$

Since $u_0 \omega^i v_0 \omega^j = u_0 v_0 \omega^{i+j} = -p/3$ only for $i + j = 3$ we conclude that

$$z_i = u_0 \omega^i + v_0 \omega^{3-i}, \quad i = 0, 1, 2,$$

are the three solutions of equation (5.13).

5.23 Fourth degree equations. Suppose we want to solve

$$P(z) = a_0 z^4 + a_1 z^3 + a_2 z^2 + a_3 z + a_4, \quad a_i \in \mathbb{C}, \quad a_0 \neq 0. \quad (5.14)$$

We observe that, if $\alpha = -a_1/4a_0$, then $P'''(\alpha) = 0$, hence

$$P(z) = a_0(z - \alpha)^4 + \frac{P''(\alpha)}{2!}(z - \alpha)^2 + P'(\alpha)(z - \alpha) + P(\alpha), \quad (5.15)$$

and z solves (5.14) if and only if $y := z - \alpha$ solves (5.15). Therefore it suffices to solve

$$z^4 + pz^2 + qz + r = 0, \quad p, q, r \in \mathbb{R}. \quad (5.16)$$

We look for solutions of the type $z = u + v + w$. Inserting into the equation we find

$$\begin{aligned} z^4 - 2(u^2 + v^2 + w^2)z^2 - 8uvwz + (u^2 + v^2 + w^2)^2 \\ - 4(u^2v^2 + v^2w^2 + u^2w^2) = 0, \end{aligned}$$

therefore $z = u + v + w$ is a solution if

$$\begin{cases} u^2 + v^2 + w^2 = -p/2, \\ uvw = -q/8, \\ u^2v^2 + v^2w^2 + u^2w^2 = (p^2 - 4r)/16. \end{cases}$$

By computation, see Exercise 5.61, u^2, v^2 and w^2 are the three solutions y_1, y_2, y_3 of the third degree equation

$$y^3 + \frac{p}{2}y^2 + \frac{p^2 - 4r}{16}y - \frac{q^2}{64} = 0.$$

Consequently u, v, w are to be chosen among the square roots $\pm u_0, \pm v_0, \pm w_0$ of y_1, y_2, y_3 . If we choose w_0 in such a way that $u_0v_0w_0 = -q/8$, then we conclude that

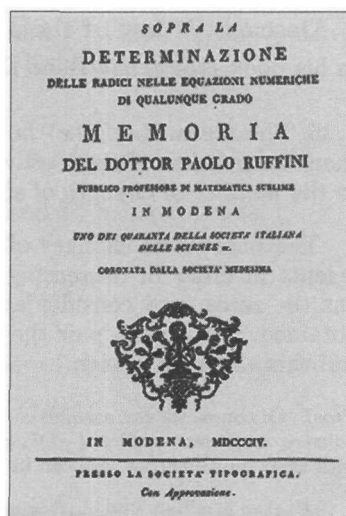
$$\begin{cases} z_1 = u_0 + v_0 + w_0, \\ z_2 = u_0 - v_0 - w_0, \\ z_3 = -u_0 + v_0 - w_0, \\ z_4 = -u_0 - v_0 + w_0 \end{cases}$$

are the four solutions of equation (5.16).

5.24 Solutions by radicals. The study of algebraic equations and, especially, the research of a procedure for solving algebraic equations, i.e., finding the roots of a given equation from its coefficients by means of a finite number of rational operations and extraction of radicals, continued till the end of the eighteenth century. In 1770 a fundamental work by Joseph-Louis Lagrange (1736–1813) appeared in the *Nouv. Mem. de l'Acad. de Berlin*. There he analyzed the methods for solving equations of degree at most 4 and set a new basis for the study of higher order equations. In 1799 Carl Friedrich Gauss (1777–1855) provided a first rigorous proof of the fundamental theorem of algebra (incomplete proofs had been given by Jean d'Alembert (1717–1783), Leonhard Euler (1707–1783), and Gauss himself). From 1799 with the treatise *Teoria generale delle equazioni in cui si dimostra impossibile la soluzione algebrica delle equazioni di grado*



Figure 5.10. Niels Henrik Abel (1802–1829) and the frontispiece of a *Memoria* by Paolo Ruffini (1765–1822).



superiore al quarto until 1813, Paolo Ruffini (1765–1822) made several attempts to prove that general equations of degree higher than four could not be solved by radicals. Finally in 1824 the *Mémoire sur les équations algébriques, où l'on démontre l'impossibilité de la résolution de l'équation générale de cinquième degré* by Niels Henrik Abel (1802–1829) appeared, where a complete proof of Ruffini's attempts was given.

The problem then became that of deciding whether a specific equation was or was not solvable by radicals, and, if not, finding more complicated formulas: the fundamental ideas in this direction are due to Evariste Galois (1811–1832) with further contributions, among others, by Enrico Betti (1823–1892), Charles Hermite (1822–1901), Leopold Kronecker (1823–1891) which led to the in some sense definitive treatise by Camille Jordan (1838–1922). But this would lead us far away from our main path.

5.2.2 Distribution of the roots of a polynomial

Let P be a polynomial of degree n with real coefficients, $P(x) = \sum_{k=0}^n a_k x^k$. Without solving the equation $P(x) = 0$ we would like to obtain information about the distribution of its roots. For instance, we would like to determine whether it has real roots, and if it does, how many; or how many positive roots it has, or how many real roots lie between given limits a and b .

a. Descartes's law of signs

In his book *Geometria* René Descartes (1596–1650) proved the following.

5.25 Theorem. *Let $P(x)$ be a real polynomial. If all its roots are real, then the number of its positive roots, counted with multiplicity, is equal to the number of changes of sign in the sequence of its coefficients.*

The number of changes of sign is defined as follows: we list all coefficients in order of decreasing power index, including a_n and a_0 , cancel out the zeros, and consider all pairs of successive numbers in the list so obtained: if in such a pair the signs of the numbers are different, then we call this a change of sign.

Proof. Of course we can assume $a_n > 0$. We start by observing that the sign of the first nonzero coefficient of P is $(-1)^p$, p being the number of positive roots of P , counted with their multiplicities. We can in fact write

$$P(x) = a_n x^n + \cdots + a_k x^k = a_n x^k (x - x_1) \cdots (x - x_p)(x - x_{p+1}) \cdots (x - x_k)$$

if P has 0 as a root of multiplicity k , x_1, x_2, \dots, x_p are the positive roots of P and x_{p+1}, \dots, x_{n-k} are the negative ones.

The proof now proceeds by induction on the degree of P . The claim is trivial for $n = 1$. Let us suppose it for polynomials of degree $n - 1$ and consider a polynomial $P(x) = \sum_{k=0}^n a_k x^k$ of degree n .

If $a_0 = 0$, then $P(x) = xQ(x)$. Since the number of positive roots and the number of changes of sign of P and Q are equal and the claim holds for Q , it holds for P .

Suppose $a_0 \neq 0$. It is clear that the number of changes of sign in $P(x)$ is equal to the analogous number for the derivative $P'(x)$, if the sign of a_0 and the last coefficient of P' coincide, or it is one more if the signs are opposite. By the remark at the beginning of the proof, in the first case the numbers of positive roots of P and P' have the same parity (are both even or odd), and in the second case they have opposite parity. On the other hand the number of positive roots of P can be either equal to the number of positive roots of P' , or one more. Therefore in both cases the difference between the changes of sign and the number of roots of P and P' is the same.

Since the number of positive roots of P' is equal to the number of changes of sign in the coefficients of P' , by inductive assumption, the claim is proved also for P . \square

5.26 Remark. Actually one could also show: if $P(x)$ has also complex roots, then the number of positive roots is equal to, or an even number less, than the number of changes in sign in the coefficients.

b. Sturm's theorem

Descartes's law of signs does not give an answer to our initial question to know the number of real roots of a given polynomial in a given interval. After many attempts, this problem was solved by Jean-Charles-François Sturm (1803–1855) in 1835. He considers the problem of localizing the sets of zeros of a polynomial disregarding multiplicities. For this problem it suffices to consider the case of polynomials with only simple roots, since for any polynomial P , $P_0 := P(x)/\text{g.c.d.}(P(x), P'(x))$ has the same set of roots of P , and P_0 and P'_0 are coprime. So we can and do assume from now on that P and P' are coprime.

We now apply Euclid's algorithm to P and P' with a slightly different notation to find the list of polynomials $\{Q_k\}$ given by

$$\begin{cases} Q_0(x) = P(x), & Q_1(x) = P'(x), \\ Q_{k-1}(x) = q_k(x)Q_k(x) - Q_{k+1}(x), & \deg Q_{k+1} < \deg Q_k, \end{cases} \quad (5.17)$$

till the last one, $Q_\ell(x)$, that is nonzero. P and P' being coprime, $Q_\ell(x)$ is a nonzero constant. The list of polynomials

$$\{Q_0(x), Q_1(x), \dots, Q_\ell(x)\}$$

is called the *Sturm's sequence* of P .

For $a \in \mathbb{R}$ we denote by $V(a)$ the number of changes of sign in the list

$$\{Q_0(a), Q_1(a), \dots, Q_\ell(a)\}$$

not counting possible zeros.

5.27 Theorem (Sturm). *The number of roots of P in $[a, b]$ is equal to $V(a) - V(b)$.*

Proof. The idea is to look at how $V(\xi)$ changes when ξ moves from a to b . Let I be the interval between two consecutive roots of the Q_i 's, $i = 0, \dots, \ell - 1$. By continuity $\text{sgn}(Q_i(\xi))$ is constant in I , from which we infer $V(\xi)$ is constant in I . Thus $V(\xi)$ is a piecewise constant function on $[a, b]$ that eventually jumps at the roots of one of the Q_i 's.

Let us now look at $V(\xi)$ when ξ , moving from a to b , passes through one of the roots of the polynomials $Q_0, \dots, Q_{\ell-1}$. First we observe that if $Q_i(\xi) = 0$, then $Q_{i-1}(\xi)$ and $Q_{i+1}(\xi)$ are nonzero, since $Q_j(\xi) = Q_{j+1}(\xi) = 0$ and (5.17) would give $Q_{j+2}(\xi) = 0, \dots, Q_\ell(\xi) = 0$: a contradiction.

(i) Assume then

$$Q_0(\gamma) = 0, \quad Q_1(\gamma) \neq 0, \quad \dots, \quad Q_{\ell-1}(\gamma) \neq 0, \quad Q_\ell(\gamma) \neq 0.$$

The continuity of the Q_i 's yields a $\delta > 0$ such that

$$\text{sgn } Q_i(\gamma - \delta) = \text{sgn } Q_i(\gamma + \delta) \quad \text{for } i = 1, \dots, \ell.$$

Since $Q_1(\gamma) \neq 0$, Q_0 is monotone in a neighborhood of γ , we may assume

$$\text{sgn } Q_0(\gamma - \delta) = -\text{sgn } Q_1(\gamma - \delta) \quad \text{sgn } Q_0(\gamma + \delta) = \text{sgn } Q_1(\gamma + \delta).$$

In this case, we therefore conclude

$$V(\gamma - \delta) - V(\gamma + \delta) = 1.$$

(ii) If instead γ is a root of

$$Q_i(x) = 0 \quad \text{for some } i = 1, 2, \dots, \ell - 1,$$

then (5.17) yields $Q_{i-1}(\gamma) = -Q_{i+1}(\gamma)$, therefore Q_{i-1} and Q_{i+1} have opposite sign in a neighborhood of γ , say $[\gamma - \delta, \gamma + \delta]$, since they are not zero. Hence the following two tables are possible

x	$\gamma - \delta$	γ	$\gamma + \delta$	x	$\gamma - \delta$	γ	$\gamma + \delta$
Q_{i-1}	+	+	+	Q_{i-1}	-	-	-
Q_i	\pm	0	\pm	Q_i	\pm	0	\pm
Q_{i+1}	-	-	-	Q_{i+1}	+	+	+

In this case we then conclude that

$$V(\gamma - \delta) - V(\gamma + \delta) = 0.$$

From (i) (ii), we infer that $V(\xi)$ jumps at γ if and only if γ is a root of Q_0 , and the jump is $+1$, thus concluding that $V(a) - V(b)$ equals the number of zeros of $Q_0 = P$. \square

5.3 Rational Functions

5.28 Definition. A rational function $R(z)$ is the quotient $R(z) = A(z)/B(z)$ of two polynomial functions $A, B : \mathbb{C} \rightarrow \mathbb{C}$. $R(z)$ is therefore defined for all $z \in \mathbb{C}$ such that $B(z) \neq 0$.

a. Decomposition in \mathbb{C}

Hermite's formula allows to decompose every rational function $A(z)/B(z)$ as the sum of a polynomial (which is nonzero if and only if $\deg A \geq \deg B$) and of *simple rational functions*, that is of functions of the type

$$\frac{\lambda}{(z - \alpha)^k}$$

where $\lambda \in \mathbb{C}$, $k \in \mathbb{N}$, and α is a root of B .

Of course we can assume that A and B are coprime. Also, if $\deg A \geq \deg B$, we may divide A by B obtaining $A(z) = B(z)Q(z) + R(z)$ with $\deg R(z) < \deg B(z)$, i.e.,

$$\frac{A(z)}{B(z)} = Q(z) + \frac{R(z)}{B(z)}, \quad \deg R < \deg B.$$

Therefore in the sequel of this section we shall always assume that A and B are coprime and $\deg A < \deg B$.

Let $\alpha_1, \alpha_2, \dots, \alpha_p$ be the roots of B with multiplicities k_1, k_2, \dots, k_p so that

$$B(z) = (z - \alpha_1)^{k_1} (z - \alpha_2)^{k_2} \cdots (z - \alpha_p)^{k_p}, \quad k_1 + k_2 + \cdots + k_p = n.$$

Fix one of the roots, denote it by α and denote its multiplicity by k , so that

$$B(z) = (z - \alpha)^k Q_\alpha(z), \quad Q_\alpha(\alpha) \neq 0,$$

while $Q_\alpha(\beta) = 0$ for any root β of B with $\beta \neq \alpha$. Notice also that, by Taylor's formula,

$$Q_\alpha(\alpha) = \frac{D^k B(\alpha)}{k!}.$$

Hermite's decomposition formula is a consequence of the following lemma, which allows us to decrease the degree of the denominator.

5.29 Lemma. *Let α be a root of B with multiplicity k , and let $Q_\alpha(z)$ be such that $B(z) = (z - \alpha)^k Q_\alpha(z)$. We have*

$$\frac{A(z)}{B(z)} = \frac{\lambda_{\alpha,k}}{(z - \alpha)^k} + \frac{R(z)}{(z - \alpha)^{k-1} Q_\alpha(z)} \quad (5.18)$$

where

$$\lambda_{\alpha,k} := \frac{A(\alpha)}{Q_\alpha(\alpha)} = \frac{k! A(\alpha)}{D^k B(\alpha)}, \quad R(z) := \frac{A(z) - \lambda_{\alpha,k} Q_\alpha(z)}{z - \alpha}.$$

Moreover $\deg R < \deg B - 1$ and $R(z)$ and Q_α are coprime.

Proof. We have

$$\frac{A(z)}{B(z)} = \frac{\lambda Q_\alpha(z)}{B(z)} + \frac{A(z) - \lambda Q_\alpha(z)}{B(z)} = \frac{\lambda}{(z - \alpha)^k} + \frac{A(z) - \lambda Q_\alpha(z)}{(z - \alpha)^k Q_\alpha(z)}.$$

Since $A(\alpha) - \lambda Q_\alpha(\alpha) = 0$ if $\lambda = \lambda_{\alpha,k}$, we have $A(z) - \lambda_{\alpha,k} Q_\alpha(z) = R(z)(z - \alpha)$, proving (5.18) and that $\deg R < \deg B - 1$. It remains to prove that R and Q_α are coprime. In fact, if β is a common root to R and Q_α , then $A(\beta) = A(\beta) - \lambda_{\alpha,k} Q_\alpha(\beta) = R(\beta)(\beta - \alpha) = 0$, hence β is a common root to A and B , a contradiction. \square

Iterating Lemma 5.29 we get the following.

5.30 Theorem. *Let A, B be coprime polynomials with $\deg A < \deg B$ and let α be a root of B with multiplicity k . Then we can find $\lambda_{\alpha,k}, \lambda_{\alpha,k-1}, \dots, \lambda_{\alpha,1} \in \mathbb{C}$ and a polynomial R_α coprime with $Q_\alpha(z)$ with $\deg R_\alpha < \deg B - k$, such that*

$$\frac{A(z)}{B(z)} = \sum_{j=1}^k \frac{\lambda_{\alpha,j}}{(z - \alpha)^j} + \frac{R_\alpha(z)}{Q_\alpha(z)}.$$

Actually $\{\lambda_{\alpha,k}\}$ and R_α can be computed as follows: Let $Q_\alpha(z)$ be such that $B(z) = (z - \alpha)^k Q_\alpha(z)$, $Q_\alpha(\alpha) \neq 0$. Set $R_{\alpha,k}(z) := A(z)$, compute iteratively for $j = k, \dots, 2, 1$,

$$\begin{cases} \lambda_{\alpha,j} := R_{\alpha,j}(\alpha)/Q_\alpha(\alpha), \\ R_{\alpha,j-1}(z) := (R_{\alpha,j} - \lambda_{\alpha,j} Q_\alpha(z))/(z - \alpha), \end{cases}$$



Figure 5.11. Charles Hermite (1822–1901).

and set $R_\alpha(z) := R_{\alpha,0}(z)$.

Finally, observe that, for any root $\beta \neq \alpha$, we have

$$R_\alpha(\beta) = \frac{A(\beta)}{(\beta - \alpha)^k}. \quad (5.19)$$

5.31 Theorem (Hermite). Let A, B be coprime polynomials in $\mathbb{C}[z]$ with $\deg A < \deg B =: n$. Let $k(\alpha)$ be the multiplicity of the root α of B . Then we can find uniquely determined complex numbers $\lambda_{\alpha,j}$, $j = 1, \dots, k(\alpha)$ such that

$$\frac{A(z)}{B(z)} = \sum_{\alpha \text{ root of } B} \sum_{j=1}^{k(\alpha)} \frac{\lambda_{\alpha,j}}{(z - \alpha)^j} \quad (5.20)$$

where the $\lambda_{\alpha,j}$ can be computed as follows: let $Q_\alpha(z)$ be such that $B(z) = (z - \alpha)^{k(\alpha)} Q_\alpha(z)$, $Q_\alpha(\alpha) \neq 0$. Set $R_{\alpha,k(\alpha)}(z) := A(z)$, then compute iteratively for $j = k(\alpha), \dots, 2, 1$

$$\begin{cases} \lambda_{\alpha,j} := R_{\alpha,j}(\alpha)/Q_\alpha(\alpha), \\ R_{\alpha,j-1}(z) := (R_{\alpha,j} - \lambda_{\alpha,j} Q_\alpha(z))/(z - \alpha). \end{cases}$$

Proof. Uniqueness. Let α be a root of B with multiplicity k_α . Multiplying (5.20) by $(z - \alpha)^{k_\alpha}$, we characterize $\lambda_{\alpha,k_\alpha}$ by

$$\lambda_{\alpha,k_\alpha} := \lim_{z \rightarrow \alpha} \frac{A(z)(z - \alpha)^{k_\alpha}}{B(z)} = \frac{A(\alpha)}{Q_\alpha(\alpha)}.$$

Proceeding iteratively, the uniqueness of the decomposition follows.

Existence. The existence of the decomposition follows applying Theorem 5.30 to the roots of $B(z)$ ordered in an arbitrary order. Moreover, because of the uniqueness, we get the same decomposition starting from an arbitrary root, hence the algorithm. \square

5.32 Remark. If the roots of B are distinct, Hermite's formula becomes particularly simple. Let A and B be coprime and $\deg A < \deg B$. Denote by $\alpha_1, \alpha_2, \dots, \alpha_n$ the roots of B , and set

$$Q_i(z) := Q_{\alpha_i}(z) = B(z)/(z - \alpha_i)$$

so that

$$Q_i(\alpha_i) = B'(\alpha_i), \quad Q_i(\alpha_j) = 0 \text{ for } j \neq i; \quad (5.21)$$

then we have

$$\frac{A(z)}{B(z)} = \sum_{j=1}^n \frac{\lambda_j}{z - \alpha_j}$$

where

$$\lambda_j := \frac{A(\alpha_j)}{Q_j(\alpha_j)} = \frac{A(\alpha_j)}{B'(\alpha_j)}.$$

A direct proof of Hermite's formula in this case can be done as follows. Write $1/(z - \alpha_j) = Q_j(z)/B(z)$ so that

$$\sum_{j=1}^n \frac{\lambda_j}{z - \alpha_j} = \frac{\sum_{j=1}^n \lambda_j Q_j(z)}{B(z)}. \quad (5.22)$$

The degree of the polynomial $C(z) := \sum_{j=1}^n \lambda_j Q_j(z)$ is less than n , moreover for all $i = 1, \dots, n$ (5.21) yields

$$C(\alpha_i) = \sum_{j=1}^n \frac{A(\alpha_j)}{Q_j(\alpha_j)} Q_j(\alpha_i) = A(\alpha_i).$$

Since C and A agree on n points, we have $C = A$ and the claim follows from (5.22).

5.33 Example. Suppose we want to decompose

$$\frac{1}{(z^2 + 1)(z - 1)^3}.$$

We start with the root 1 with multiplicity 3. In the algorithm of Theorem 5.30 $A(z) = 1$, $Q_1(z) = z^2 + 1$ and $R_3(z) = A(z) = 1$. Then we compute

$$\lambda_{1,3} = \frac{A(1)}{Q_1(1)} = \frac{1}{2},$$

$$R_2(z) = (R_3(z) - \frac{1}{2}Q_1(z)) : (z - 1)$$

$$= (1 - \frac{1}{2}(z^2 + 1)) : (z - 1) = -\frac{1}{2}(z^2 - 1) : (z - 1) = -\frac{1}{2}(z + 1),$$

$$\lambda_{1,2} = R_2(1)/Q_1(1) = -\frac{1}{2},$$

$$R_1(z) = (R_2(z) + \frac{1}{2}(z^2 + 1)) : (z - 1)$$

$$= (-\frac{1}{2}(z + 1) + \frac{1}{2}(z^2 + 1)) : (z - 1) = \frac{1}{2}z(z - 1) : (z - 1) = \frac{1}{2}z,$$

$$\lambda_{1,1} = R_1(1)/Q_1(1) = \frac{1}{4},$$

$$R(z) = R_0(z) = (R_1(z) - \frac{1}{4}(z^2 + 1)) : (z - 1) = -\frac{1}{4}(z - 1)^2 : (z - 1) = -\frac{1}{4}(z - 1),$$

finding

$$\frac{1}{(z^2+1)(z-1)^3} = \frac{1}{2} \frac{1}{(z-1)^3} - \frac{1}{2} \frac{1}{(z-1)^2} + \frac{1}{4} \frac{1}{z-1} - \frac{z-1}{4(z^2+1)}.$$

It remains to decompose

$$-\frac{z-1}{4(z^2+1)}.$$

As $z^2+1 = (z-i)(z+i)$, the roots of the denominator are i and $-i$ with multiplicity 1. Therefore it suffices to compute

$$\lambda_{i,1} = \frac{-(z-1)/4}{z+i} \text{ for } z=i, \quad \lambda_{-i,1} = \frac{-(z-1)/4}{z-i} \text{ for } z=-i,$$

that is $\lambda_{i,1} = -\frac{1}{8}(1+i)$, $\lambda_{-i,1} = -\frac{1}{8}(1-i)$, to conclude

$$\frac{1}{(z^2+1)(z-1)^3} = \frac{1}{2} \frac{1}{(z-1)^3} - \frac{1}{2} \frac{1}{(z-1)^2} + \frac{1}{4} \frac{1}{z-1} - \frac{1}{8} \frac{1+i}{z-i} - \frac{1}{8} \frac{1-i}{z+i}.$$

Alternatively, we can proceed as follows. From Hermite's rule we know the existence and uniqueness of a decomposition

$$\frac{1}{(z^2+1)(z-1)^3} = \frac{a}{(z-1)^3} + \frac{b}{(z-1)^2} + \frac{c}{z-1} + \frac{d}{z-i} + \frac{e}{z+i}$$

for suitable $a, b, c, d, e \in \mathbb{C}$. We can compute those coefficients by reducing to the common denominator. This way the polynomials in the numerators have to be equal, hence their coefficients have to be equal, by the principle of identity of polynomials. This yields a system of five linear equations in a, b, c, d, e that, once solved, yields the values of a, b, c, d, e .

b. Decomposition in \mathbb{R}

If $A(x), B(x) \in \mathbb{R}[x]$ are two polynomials with real coefficients, one can decompose $A(x)/B(x)$ as a sum of a polynomial (that is not zero if and only if $\deg A \geq \deg B$) and of simple rational functions with *real* coefficients. In fact, recalling that nonreal roots of B come in couples of conjugate complex numbers, by the complex Hermite's formula, Theorem 5.31, we infer

$$\begin{aligned} \frac{A(z)}{B(z)} &= \sum_{\substack{\alpha \text{ root of } B \\ \alpha \in \mathbb{R}}} \sum_{j=1}^{k(\alpha)} \frac{\lambda_{\alpha,j}}{(z-\alpha)^j} \\ &+ \sum_{\substack{\alpha \text{ root of } B \\ \Im(\alpha) > 0}} \sum_{j=1}^{k(\alpha)} \frac{\lambda_{\alpha,j}}{(x-\alpha)^j} + \sum_{\substack{\alpha \text{ root of } B \\ \Im(\alpha) > 0}} \sum_{j=1}^{k(\alpha)} \frac{\lambda_{\bar{\alpha},j}}{(x-\bar{\alpha})^j} \end{aligned} \quad (5.23)$$

for all $z \in \mathbb{C}$ with $B(z) \neq 0$. Going through the iterative scheme in Theorem 5.31 which yields the $\lambda_{\alpha,j}$, taking into account that A and B have real coefficients, we infer

$$Q_{\bar{\alpha}}(\bar{z}) = \overline{Q_{\alpha}(z)}, \quad R_{\bar{\alpha},j}(\bar{z}) = \overline{R_{\alpha,j}(z)}, \quad \lambda_{\bar{\alpha},j} = \overline{\lambda_{\alpha,j}},$$

hence from (5.23) the *Hermite decomposition formula in \mathbb{R}*

$$\frac{A(x)}{B(x)} = \sum_{\substack{\alpha \text{ root of } B \\ \alpha \in \mathbb{R}}} \sum_{j=1}^{k(\alpha)} \frac{\lambda_{\alpha,j}}{(x-\alpha)^j} + 2 \sum_{\substack{\alpha \text{ root of } B \\ \Im(\alpha) > 0}} \sum_{j=1}^{k(\alpha)} \Re\left(\frac{\lambda_{\alpha,j}}{(x-\alpha)^j}\right) \quad (5.24)$$

for $x \in \mathbb{R}$ with $B(x) \neq 0$.

Notice that if all roots of B are simple, formula (5.24) reduces to

$$\frac{A(x)}{B(x)} = \sum_{\substack{\alpha \text{ root of } B \\ \alpha \in \mathbb{R}}} \frac{\lambda_\alpha}{x-\alpha} + \sum_{\substack{\alpha \text{ root of } B \\ \Im(\alpha) > 0}} \left(\frac{\lambda_\alpha}{x-\alpha} + \frac{\overline{\lambda_\alpha}}{x-\overline{\alpha}} \right)$$

where $\lambda_\alpha := \frac{A(\alpha)}{B'(\alpha)}$ for every root α of B . Therefore

5.34 Corollary. *Let $A(x)$ and $B(x) \in \mathbb{R}[x]$ be coprime real polynomials with $\deg A < \deg B$. Suppose that $B(x)$ has only simple roots. Then*

$$\frac{A(x)}{B(x)} = \sum_{\substack{\alpha \text{ root of } B \\ \alpha \in \mathbb{R}}} \frac{\lambda_\alpha}{x-\alpha} + \sum_{\substack{\alpha \text{ root of } B \\ \Im(\alpha) > 0}} \frac{\ell_\alpha(x-p_\alpha) + m_\alpha}{x^2 - 2p_\alpha x + q_\alpha},$$

where $\lambda_\alpha := \frac{A(\alpha)}{B'(\alpha)}$ for every root $\alpha \in \mathbb{C}$ of B and, if $\Im(\alpha) > 0$,

$$p_\alpha := \Re(\alpha), \quad q_\alpha := |\alpha|^2, \quad \ell_\alpha := 2\Re(\lambda_\alpha), \quad m_\alpha := -2\Im(\lambda_\alpha)\Im(\alpha).$$

c. Integration of rational functions

Of course, Hermite's decomposition allows us to integrate and express the indefinite integral of every rational function in terms of elementary functions. Here we just state the following.

5.35 Proposition. *Under the assumptions and with the notation of Corollary 5.34, we have*

$$\begin{aligned} \int \frac{A(x)}{B(x)} dx &= \sum_{\substack{\alpha \text{ root of } B \\ \alpha \in \mathbb{R}}} \lambda_\alpha \log|x-\alpha| \\ &+ \sum_{\substack{\alpha \text{ root of } B \\ \Im(\alpha) > 0}} \left(\Re(\lambda_\alpha) \log(x^2 - 2p_\alpha x + q_\alpha) - 2\Im(\lambda_\alpha) \arctan\left(\frac{x - \Re(\alpha)}{\Im(\alpha)}\right) \right). \end{aligned}$$

5.36 Example. Let us compute

$$\int_{-\infty}^{+\infty} \frac{x^{2m}}{1+x^{2n}} dx, \quad m, n \in \mathbb{N}, \quad m < n.$$

The roots of $B(x) = 1 + x^{2n}$ are the $2n$ -th roots of -1 ,

$$\beta_j := \exp\left(i\pi \frac{2j+1}{2n}\right), \quad j = 0, \dots, 2n-1.$$

The roots with positive imaginary part are β_j , $j = 0, \dots, n-1$. According to Proposition 5.35, we need to compute the numbers $\mu_j := x^{2m}/D(1+x^{2n})$ at $x = \beta_j$. Since $\beta_j^{2n} = -1$ we get

$$\mu_j := \frac{\beta_j^{2m}}{2n\beta_j^{2n-1}} = -\frac{1}{2n}\beta_j^{2m+1} = -\frac{1}{2n}\exp(i\alpha(2j+1))$$

where

$$\alpha := \frac{2m+1}{2n}\pi.$$

We remark that

$$\alpha(2j+1) + \alpha 2(n-1-j) + 1 = \pi(2m+1),$$

hence

$$\arg(\mu_j) + \arg(\mu_{n-1-j}) = \pi.$$

Also, since $|\mu_j| = 1 \forall j$, we have $\Re(\mu_j) = -\Re(\mu_{n-1-j})$. Finally, taking into account that $p_{n-j-1} = \Re(\beta_{n-1-j}) = -\Re(\beta_j) = -p_j$, Proposition 5.35 yields

$$\begin{aligned} & \int_{-\infty}^{+\infty} \frac{x^{2m}}{1+x^{2n}} dx \\ &= \sum_{j=0}^{(n-1)/2} \Re(\mu_j) \log \left(\frac{x^2 - 2p_j x + 1}{x^2 + 2p_j x + 1} \right) \Big|_{-\infty}^{+\infty} - 2 \sum_{j=0}^{n-1} \Im(\mu_j) \arctan \left(\frac{x - p_j}{\Im(\beta_j)} \right) \Big|_{-\infty}^{+\infty} \\ &= -2\pi \sum_{j=0}^{n-1} \Im(\mu_j) = -2\pi \Im \left(\sum_{j=0}^{n-1} \mu_j \right). \end{aligned} \quad (5.25)$$

It remains to compute $\sum_{j=0}^{n-1} \mu_j$. Set $k := \exp(i\alpha)$ and $q := \exp(i2\alpha)$, so that

$$\sum_{j=0}^{n-1} \mu_j := \frac{k}{2n} \sum_{j=0}^{n-1} q^j = \frac{1}{2n} \frac{k(1-q^n)}{1-q}.$$

Since $q^n = \exp(i\pi 2 \frac{2m+1}{2n} n) = \exp(i\pi(2m+1)) = \exp(i\pi) = -1$, we deduce

$$\sum_{j=0}^{n-1} \mu_j = \frac{1}{2n} \frac{2e^{i\alpha}}{1-e^{2i\alpha}} = \frac{1}{2n} \frac{-i}{\frac{e^{i\alpha} - e^{-i\alpha}}{2i}} = \frac{1}{2n} \frac{-i}{\sin \alpha}. \quad (5.26)$$

Therefore from (5.25) and (5.26) we conclude that, for $n, m \in \mathbb{N}$, $m < n$, we have

$$\int_{-\infty}^{+\infty} \frac{x^{2m}}{1+x^{2n}} dx = \frac{\pi}{n} \frac{1}{\sin \alpha} \approx \frac{\pi}{n} \frac{1}{\sin \left(\frac{2m+1}{2n} \pi \right)}. \quad (5.27)$$

5.37 ¶¶. Following the same path of the Hermite formula for complex rational functions, prove the following.

Lemma. Let A, B be two polynomials with real coefficients which are coprime in $\mathbb{R}[x]$ with $\deg A < \deg B$, and let α be a nonreal root of B of multiplicity k . Then

$$\frac{A(z)}{B(z)} = \frac{a(z-p)+b}{(z^2-2pz+q)^k} + \frac{R(z)}{(z^2-2pz+q)^{k-1}Q(z)}$$

where, if $\lambda := A(\alpha)/Q(\alpha)$, then $a = \frac{\Re(\lambda)}{\Im(\alpha)}$, $b = \Re(\lambda)$, and

$$R(z) := \frac{A(z) - (a(z-p)+b)Q(z)}{z^2-2pz+q}$$

is a real polynomial with $\deg R < \deg B - 2$.

Iterating the previous lemma over k and all roots of B , show the following.

Theorem. Let A, B be two polynomials with real coefficients which are coprime in $\mathbb{R}[x]$ with $\deg A < \deg B$. Assume that B has no real root and denote by $k(\alpha)$ the multiplicity of the root α of B . Then

$$\frac{A(z)}{B(z)} = \sum_{\substack{\alpha \text{ root of } B \\ \Im(\alpha) > 0}} \sum_{j=1}^{k(\alpha)} \frac{a_{\alpha,j}(z - p_{\alpha}) + b_{\alpha,j}}{(z^2 - 2p_{\alpha}z + q_{\alpha})^j}$$

where $p_{\alpha} = \Re(\alpha)$, $q_{\alpha} = |\alpha|^2$, and the $a_{\alpha,j}$ and $b_{\alpha,j}$ can be computed by the following procedure: Set $Q_{\alpha}(z) = B(z)/(z^2 - 2p_{\alpha}z + q_{\alpha})^k$. Set $R_{\alpha,k}(z) := A(z)$ and compute for $j = k, k-1, \dots, 1$,

- $\lambda_{\alpha,j} := R_{\alpha,j}(\beta)/Q_{\alpha}(\alpha)$,
- $a_{\alpha,j} := \Im(\lambda_{\alpha,j})/\Im(\alpha)$, $b_{\alpha,j} = \Re(\lambda_{\alpha,j})$,
- $R_{\alpha,j-1}(z) := (R_{\alpha,j}(z) - (a_{\alpha,j}(z - p_{\alpha}) + b_{\alpha,j})Q_{\alpha}(z))/ (z^2 - 2p_{\alpha}z + q_{\alpha})$.

5.38 ¶. Prove the following.

Theorem. Let A, B be two polynomials with real coefficients which are coprime in $\mathbb{R}[x]$ with $\deg A < \deg B$. Let $N := \text{g.c.d.}(B, B')$ and $S := B/N$. Then there exist polynomials M and R with real coefficients with $\deg M < \deg N$, $\deg R < \deg S$ such that

$$\frac{A(x)}{B(x)} = \frac{d}{dx} \left(\frac{M(x)}{N(x)} \right) + \frac{R(x)}{S(x)}.$$

[Hint: Use Hermite's formula (5.23).]

5.4 Sinusoidal Functions and Their Sums

The existence in nature of periodic phenomena, i.e., phenomena that recur after some time, attracted the attention and the interest of man and probably was one of the main starting points for organized knowledge. Next to the totally fortuitous events of life, there stood out a number of more or less regular phenomena that were predictable. Seasons, the apparent motion of the moon, of the sun, of fixed stars and planets could be predicted by everyone. On the other hand, other phenomena such as solar and lunar eclipses could be predicted only by a few people, usually the high priests.

With the creation of modern science and the systematic use of calculus to investigate reality, several periodic phenomena were studied in detail (oscillations, vibrations, waves) and, once more, the ondulatory model became relevant in the nineteenth century in order to understand the nature of light and electromagnetic radiation.

5.4.1 Trigonometric polynomials

An elementary periodic phenomenon is clearly the uniform circular motion, described as we know (compare, e.g., [GM1]) by the functions sine and cosine.

5.39 Definition. A sinusoidal signal, or a circular or harmonic function of pulse or angular frequency ω is a solution of the equation of simple harmonic motion $x''(t) + \omega^2 x(t) = 0$.

All harmonic functions with pulse $\omega \neq 0$ have the form

$$a \cos \omega t + b \sin \omega t, \quad a, b \in \mathbb{R},$$

compare [GM1], or, if we set $A := \sqrt{a^2 + b^2}$ and φ is such that $\cos \varphi = a/\sqrt{a^2 + b^2}$, $\sin \varphi := -b/\sqrt{a^2 + b^2}$, all functions of the form

$$x(t) = A \cos(\omega t + \varphi), \quad A \geq 0.$$

A is the *amplitude* and φ the *phase* at time $t = 0$ of the circular motion $x(t)$. As we have seen, complex notation yields simpler formulas, since, by Euler's formulas

$$x(t) = Ae^{i\varphi} e^{i\omega t}.$$

The phase is of course defined modulo an integer multiple of 2π , therefore, if we want to have a definite number, we need to fix a determination of the angle.

a. Periodic functions

5.40 Definition. A function $f : \mathbb{R} \rightarrow \mathbb{R}$ is said to be periodic of period T if $f(t+T) = f(t) \forall t \in \mathbb{R}$.

If f and g are periodic with period T , then $f+g$ and fg are periodic of the same period T ; moreover if f is differentiable, then f' is also periodic of period T .

5.41 ¶. In general, the primitive of a periodic function is not periodic, for instance, $1 + \cos t$ is 2π -periodic, but $t + \sin t$ is not. Show the following.

Proposition. Let f be a periodic, continuous function with period T . Then $\int_0^x f(t) dt$ is periodic of period T if and only if f has integral mean zero over a period, $\int_0^T f(t) dt = 0$.

If f is a periodic function of period T , then f is also periodic with periods kT , $k \in \mathbb{N}$, $k \geq 1$. A sinusoidal signal with pulse ω , $f(t) := A \cos(\omega t + \varphi)$ has a minimum period $T := 2\pi/\omega$, and then it is periodic with all the periods $\frac{2\pi}{\omega} k$, $k \in \mathbb{N}$, $k \geq 1$. Notice however that not every periodic function has a minimum period. For instance the *Dirichlet function*

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

is periodic with period q for all $q \in \mathbb{Q}_+$. The “minimum period” should then be zero, but this is meaningless.

b. Trigonometric polynomials

As we shall see in Theorem 5.55, the sum of sinusoidal signals of pulses ω_1 and ω_2 is periodic if and only if ω_1/ω_2 is rational, that is, if both pulses are integer multiples of a fundamental pulse.

5.42 Definition. A trigonometric polynomial of degree n and period T is a periodic function of the form

$$P(t) = \frac{a_0}{2} + \sum_{k=1}^n \left(a_k \cos\left(\frac{2\pi}{T}kt\right) + b_k \sin\left(\frac{2\pi}{T}kt\right) \right), \quad t \in \mathbb{R}, \quad (5.28)$$

with $a_k, b_k \in \mathbb{R}$.

Notice that the quotients of the frequencies k/T of the components are rational. The terminology that follows comes from acoustics. The human ear considers as *harmonic* the sounds that have rational quotients of pulses and, actually, with quotient p/q with p, q small. Then

- (i) $a_0/2$, or more precisely the constant function $x \rightarrow a_0/2$, is the *continuous component* of P ,
- (ii) the function $t \rightarrow a_1 \cos\left(\frac{2\pi}{T}t\right) + b_1 \sin\left(\frac{2\pi}{T}t\right)$ is the *fundamental harmonic* of P ; the function $t \rightarrow a_k \cos\left(\frac{2\pi}{T}kt\right) + b_k \sin\left(\frac{2\pi}{T}kt\right)$, $k \geq 2$, is the k -th harmonic of P .

If we set

$$A_k := \begin{cases} a_0/2 & \text{if } k = 0, \\ \sqrt{a_k^2 + b_k^2} & \text{if } k = 1, \dots, n \end{cases}$$

and φ_k is such that

$$\cos \varphi_k = a_k / \sqrt{a_k^2 + b_k^2}, \quad \sin \varphi_k = -b_k / \sqrt{a_k^2 + b_k^2},$$

we can write P as

$$P(t) = A_0 + \sum_{k=1}^n A_k \cos\left(\frac{2\pi}{T}kt + \varphi_k\right).$$

The lists $\{A_k\}$, $k = 0, \dots, n$, and $\{\varphi_k\}$, $k = 1, \dots, n$, are called respectively the *amplitude* and the *phase spectrum* of P .

When dealing with trigonometric polynomials, the complex notation is useful to write formulas which are handier to manipulate. In fact, taking into account Euler's formulas, we can write $P(t)$ in (5.28) as

$$P(t) = \sum_{k=-n}^n c_k e^{i \frac{2\pi}{T} kt}, \quad t \in \mathbb{R} \quad (5.29)$$

where

$$c_k = \begin{cases} \frac{a_k - ib_k}{2} = A_k e^{-i\varphi_k} & \text{if } k \geq 1, \\ \frac{a_0}{2} = A_0 & \text{if } k = 0, \\ \overline{c_{-k}} & \text{if } k \leq -1. \end{cases}$$

Observe that, while each term is complex, the sum

$$c_{-k} e^{-i \frac{2\pi}{T} kt} + c_k e^{i \frac{2\pi}{T} kt} = 2\Re\left(c_k e^{i \frac{2\pi}{T} kt}\right)$$

is real for each $k \in \{-n, \dots, n\}$.

More generally, we set the following.

5.43 Definition. A complex trigonometric polynomial of degree n and period T is a periodic function with complex values, $P : \mathbb{R} \rightarrow \mathbb{C}$, of the type

$$P(t) = \sum_{k=-n}^n c_k \exp\left(i \frac{2\pi}{T} kt\right), \quad t \in \mathbb{R}, \quad (5.30)$$

where $c_k \in \mathbb{C}$ for $k = -n, \dots, n$. The vector $\{c_k\}_{-n \leq k \leq n} \in \mathbb{C}^{2n+1}$ is called the complex spectrum, or simply the spectrum of P .

The class of all complex trigonometric polynomials is denoted by $\mathcal{P}_{n,T}$.

Notice that $f + g$ and $\lambda f \in \mathcal{P}_{n,T}$ if $f, g \in \mathcal{P}_{n,T}$ and $\lambda \in \mathbb{C}$, or, as we say, $\mathcal{P}_{n,T}$ is a complex vector space. Finally, notice that $P(\frac{T}{2\pi}t) \in \mathcal{P}_{n,2\pi}$ if $P \in \mathcal{P}_{n,T}$.

c. Spectrum and energy identity

By definition, a complex trigonometrical polynomial is defined uniquely by its spectrum, and the surjective map $\Phi : \mathbb{C}^{2n+1} \rightarrow \mathcal{P}_{n,T}$ given by

$$(c_{-n}, \dots, c_n) \longrightarrow f(t) := \sum_{k=-n}^n c_k \exp\left(i \frac{2\pi}{T} kt\right) \quad (5.31)$$

is linear. Thus $\mathcal{P}_{n,T}$ is a vector space of dimension less than or equal to $2n + 1$. A relevant fact is that one can compute the complex amplitudes of the harmonics from the sum of the harmonics, thus proving that Φ is injective and therefore that the dimension of $\mathcal{P}_{n,T}$ is $2n + 1$.

5.44 Proposition. Let $P(t) = \sum_{k=-n}^n c_k e^{i \frac{2\pi}{T} kt} \in \mathcal{P}_{n,T}$.

(i) For all $k = -n, \dots, n$ we have

$$c_k = \frac{1}{T} \int_0^T P(t) e^{-i \frac{2\pi}{T} kt} dt \quad (5.32)$$

Consequently $\mathcal{P}_{n,T}$ has complex dimension $2n + 1$.

(ii) The energy equality holds

$$\frac{1}{T} \int_0^T |P(t)|^2 dt = \sum_{k=-n}^n |c_k|^2. \quad (5.33)$$

The proof of Proposition 5.44 is a simple consequence of the following computation.

5.45 Lemma. Let $k \in \mathbb{Z}$. Then

$$\frac{1}{T} \int_0^T e^{i \frac{2\pi}{T} kt} dt = \begin{cases} 0 & \text{if } k \neq 0, \\ 1 & \text{if } k = 0. \end{cases}$$

Proof. If $k = 0$, we have $e^{ikt} = 1$, hence $\frac{1}{T} \int_0^T e^{i \frac{2\pi}{T} kt} dt = 1$. If $k \neq 0$, writing $e^{i \frac{2\pi}{T} kt} = \cos\left(\frac{2\pi}{T} kt\right) + i \sin\left(\frac{2\pi}{T} kt\right)$ and noticing that $\cos \theta$ and $\sin \theta$ have zero mean average over an interval of size 2π , we conclude that $\frac{1}{T} \int_0^T e^{i \frac{2\pi}{T} kt} dt = 0$. \square

Introducing the *Kronecker symbol* δ_{hk} ,

$$\delta_{hk} = \begin{cases} 1 & \text{if } h = k, \\ 0 & \text{if } h \neq k, \end{cases}$$

Lemma 5.45 yields

$$\frac{1}{T} \int_0^T e^{i \frac{2\pi}{T} (h-k)t} dt = \delta_{hk} \quad \forall h, k \in \mathbb{Z}. \quad (5.34)$$

Proof of Proposition 5.44. (i) From (5.34)

$$\begin{aligned} \frac{1}{T} \int_0^T P(t) e^{-i \frac{2\pi}{T} kt} dt &= \frac{1}{T} \int_0^T \sum_{h=-n}^n c_h e^{i \frac{2\pi}{T} ht} e^{-i \frac{2\pi}{T} kt} dt \\ &= \sum_{h=-n}^n c_h \left(\frac{1}{T} \int_0^T e^{i \frac{2\pi}{T} (h-k)t} dt \right) = \sum_{h=-n}^n c_h \delta_{hk} = c_k. \end{aligned}$$

(ii) We have

$$|P(t)|^2 = \left(\sum_{k=-n}^n c_k e^{i \frac{2\pi}{T} kt} \right) \overline{\left(\sum_{h=-n}^n c_h e^{i \frac{2\pi}{T} ht} \right)} = \sum_{h,k=-n,n} c_k \overline{c_h} e^{i \frac{2\pi}{T} (k-h)t}.$$

Thus we conclude from (5.34) that

$$\frac{1}{T} \int_0^T |P(t)|^2 dt = \sum_{h,k=-n,n} c_k \overline{c_h} \delta_{kh} = \sum_{k=-n}^n c_k \overline{c_k} = \sum_{k=-n}^n |c_k|^2.$$

□

5.46 Remark. Let $f \in C^0(\mathbb{R})$ be periodic with period T . The *energy* of f is defined to be

$$\frac{1}{T} \int_0^T |f(t)|^2 dt.$$

Of course the energy of the continuous component of P is $|c_0|^2$, while the energy of the k -th harmonic of $P(t)$ is

$$\frac{1}{T} \int_0^T |c_k e^{i \frac{2\pi}{T} kt} + c_{-k} e^{-i \frac{2\pi}{T} kt}|^2 dt = |c_k|^2 + |c_{-k}|^2.$$

The energy identity can therefore be restated as: *the energy of a trigonometric polynomial is the sum of the energies of its components.*

d. Sampling

Let $P(t) = \sum_{k=-n}^n c_k e^{i \frac{2\pi}{T} kt} \in \mathcal{P}_{n,T}$ be a complex trigonometric polynomial of order n . Trivially $P(t) = R(\exp(i \frac{2\pi}{T} t))$, where R is the rational function

$$R(z) := \sum_{k=-n}^n c_k z^k.$$

Observing that $R(z) = N(z)/z^n$ where N is a polynomial of degree $2n$, and taking into account the principle of identity of polynomials, we infer the following.

5.47 Proposition. Let $P, Q \in \mathcal{P}_{n,T}$. Suppose that P and Q agree on $2n+1$ distinct points in $[0, T[$. Then $P(t) = Q(t)$ for all $t \in \mathbb{R}$.

Not only is this true, but there is an *interpolation formula* that permits to reconstruct $P(t)$ from its values on a suitable choice of $2n+1$ points in $[0, T[$. This is an easy version of the sampling theorem of Claude Shannon (1916–2001). In order to show that formula, we introduce *Dirichlet's kernel* of order n as the trigonometric polynomial in $\mathcal{P}_{n,2\pi}$ defined by

$$D_n(t) := \sum_{k=-n}^n e^{ikt} = 1 + 2 \sum_{k=1}^n \cos kt, \quad t \in \mathbb{R}. \quad (5.35)$$

5.48 Proposition. We have

- (i) $D_n(t)$ is an even function $D_n(-t) = D_n(t)$,
- (ii) $D_n(0) = 2n + 1$ and $D_n(\pi) = (-1)^n$,
- (iii) for $t \neq 2k\pi$, $k \in \mathbb{Z}$, we have

$$D_n(t) = \frac{\sin((n + 1/2)t)}{\sin t/2},$$

- (iv) $D_n(t)$ vanishes at $t_j = \frac{2\pi}{2n+1}j$ if $j \in \mathbb{Z}$ and j is not a multiple of $2n + 1$. In particular, if $j \in [-2n, 2n]$, then

$$D_n\left(\frac{2\pi}{2n+1}j\right) = \begin{cases} 0 & \text{if } j \neq 0, \\ 2n + 1 & \text{if } j = 0. \end{cases}$$

Proof. (i), (ii), and (iv) are trivial. (iii) can be proved by induction or even directly. In fact, on account of

$$\sum_{k=0}^p z^k = 1 + z + z^2 + \cdots + z^p = \frac{1 - z^{p+1}}{1 - z}, \quad z \neq 1,$$

one computes

$$D_n(t) = \frac{1 - z^{2n+2}}{z^n(1 - z)}, \quad z := e^{it},$$

hence

$$D_n(t) = e^{-int} \frac{1 - e^{i(2n+1)t}}{1 - e^{it}} = \cdots = \frac{\sin((n + 1/2)t)}{\sin(t/2)}. \quad (5.36)$$

□

5.49 Theorem. Let $Q(x) \in \mathcal{P}_{n,T}$. Set $P(t) := Q(\frac{T}{2\pi}t) \in \mathcal{P}_{n,2\pi}$. Then

$$P(t) = \frac{1}{2n+1} \sum_{j=-n}^n P(t_j) D_n(t - t_j) \quad \forall t \in \mathbb{R},$$

where $t_j := \frac{2\pi}{2n+1}j$, $j \in \{-n, \dots, n\}$.

In other words, we can reconstruct $P(t)$ from the values $P(t_j)$ of P at t_j .

Proof. Since the formula is linear with respect to P , it is enough to prove it only for the functions e^{ihx} , $h = -n, \dots, n$, that form a basis for $\mathcal{P}_{n,2\pi}$. From the definitions of D_n and of $\{t_j\}$, we deduce

$$\begin{aligned}
\sum_{j=-n}^n e^{ihx_j} D_n(x - x_j) &= \sum_{j=-n}^n \sum_{k=-n}^n e^{ihx_j} e^{ik(x-x_j)} \\
&= \sum_{j=-n}^n \sum_{k=-n}^n e^{ikx} e^{i(h-k)x_j} = \sum_{k=-n}^n e^{ikx} \sum_{j=-n}^n e^{i(h-k)x_j} \\
&= \sum_{k=-n}^n e^{ikx} D_n\left(\frac{2\pi}{2n+1}(h-k)\right).
\end{aligned}$$

Since $h - k \in [-2n, 2n]$, (iv) of Proposition 5.48 yields

$$D_n\left(\frac{2\pi}{2n+1}(h-k)\right) = (2n+1)\delta_{hk}$$

hence

$$\sum_{k=-n}^n e^{ikx} D_n\left(\frac{2\pi}{2n+1}(h-k)\right) = (2n+1)e^{ihx}.$$

□

5.50 Example (Euler's formula). Let us prove that

$$\int_0^\infty \frac{\sin t}{t} dt = \frac{\pi}{2}. \quad (5.37)$$

We recall that the integral in (5.37) exists as an *improper integral*,

$$\int_0^\infty \frac{\sin x}{x} dx := \lim_{y \rightarrow \infty} \int_0^y \frac{\sin x}{x} dx;$$

see, e.g., Example 4.81 of [GM1]. Therefore it suffices to compute the limit of $\int_0^{x_n} \frac{\sin t}{t} dt$ as $n \rightarrow \infty$ where $\{x_n\}$ is a sequence that diverges to $+\infty$.

From (5.35) and (5.36) we infer

$$\frac{\sin(2n+1)t}{\sin t} = 1 + 2 \sum_{k=1}^n \cos 2kt$$

and, integrating over $[0, \pi/2]$,

$$\int_0^{\pi/2} \frac{\sin(2n+1)t}{\sin t} dt = \frac{\pi}{2} + 2 \sum_{k=1}^n \frac{\sin(2kt)}{2k} \Big|_0^{\pi/2} = \frac{\pi}{2}.$$

Therefore

$$\begin{aligned}
\frac{\pi}{2} - \int_0^{(2n+1)\frac{\pi}{2}} \frac{\sin t}{t} dt &= \int_0^{\pi/2} \frac{\sin((2n+1)t)}{\sin t} dt - \int_0^{\pi/2} \frac{\sin(2n+1)t}{t} dt \\
&= \int_0^{\pi/2} \frac{t - \sin t}{t \sin t} \sin(2n+1)t dt \\
&= \int_0^{\pi/2} f(t) \sin(2n+1)t dt
\end{aligned} \quad (5.38)$$

where $f(t) = \frac{t - \sin t}{t \sin t}$. Since $f(t)$ is of class C^1 in $[0, \pi/2]$, we can integrate by parts the last integral in (5.38) finding

$$\int_0^{\pi/2} f(t) \sin(2n+1)t dt = -f(t) \frac{\cos(2n+1)t}{2n+1} \Big|_0^{\pi/2} + \int_0^{\pi/2} f'(t) \frac{\cos(2n+1)t}{2n+1} dt \rightarrow 0$$

as $n \rightarrow \infty$, i.e.,

$$\int_0^{(2n+1)\frac{\pi}{2}} \frac{\sin t}{t} dt \rightarrow \frac{\pi}{2}.$$

5.4.2 Sums of sinusoidal functions

5.51 Definition. A finite sum of complex sinusoidal functions is a function $f : \mathbb{R} \rightarrow \mathbb{C}$ of the form

$$f(t) = \sum_{j=1}^n c_j \exp(i\omega_j t), \quad t \in \mathbb{R},$$

where $c_j \in \mathbb{C}$, $\omega_j \in \mathbb{R}$, $\omega_i \neq \omega_j$ for $i \neq j$ and $j = 1, \dots, n$. Each addend is referred to as a component of f , c_j is the amplitude of the component and the vector $\hat{f} := (c_1, \dots, c_n) \in \mathbb{C}^n$ is the spectrum of f .

The sum of complex sinusoidal functions identifies again the coefficients of its components. In fact, we have the following.

5.52 Theorem. Let $f : \mathbb{R} \rightarrow \mathbb{C}$ be the finite sum of complex harmonic functions

$$f(t) = \sum_{j=1}^n c_j \exp(i\omega_j t).$$

Then for any $\omega \in \mathbb{R}$,

$$\lim_{N \rightarrow +\infty} \frac{1}{2N} \int_{-N}^N f(t) \exp(-i\omega t) dt = \begin{cases} c_j & \text{if } \omega = \omega_j \text{ for some } j = 1, \dots, n, \\ 0 & \text{otherwise.} \end{cases}$$

The claim in Theorem 5.52 follows easily from the following

5.53 Lemma. For all $\omega \in \mathbb{R}$ we have

$$\lim_{N \rightarrow +\infty} \frac{1}{2N} \int_{-N}^N \exp(i\omega t) dt = \begin{cases} 0 & \text{if } \omega \neq 0, \\ 1 & \text{if } \omega = 0. \end{cases}$$

Proof. If $\omega = 0$ the claim is trivial. Suppose $\omega \neq 0$, let $T := 2\pi/\omega$ and let k be the largest integer such that $kT < N$, so that $N - kT < T$. Since $\exp(i\omega t)$ has zero average over one period interval we deduce

$$\int_{-kT}^{kT} \exp(i\omega t) dt = 0,$$

hence

$$\begin{aligned} \left| \int_{-N}^N \exp(i\omega t) dt \right| &= \left| \int_{-N}^{-kT} \exp(i\omega t) dt + \int_{kT}^N \exp(i\omega t) dt \right| \\ &\leq \int_{-N}^{-kT} |\exp(i\omega t)| dt + \int_{kT}^N |\exp(i\omega t)| dt \\ &\leq 2T. \end{aligned}$$

Therefore we conclude $\left| \frac{1}{2N} \int_{-N}^N \exp(i\omega t) dt \right| \leq \frac{T}{N} \rightarrow 0$ as $N \rightarrow +\infty$. \square

Proof of Theorem 5.52. Since

$$\frac{1}{2N} \int_{-N}^N f(t) \exp(-i\omega t) dt = \sum_{j=1}^n c_j \frac{1}{2N} \int_{-N}^N \exp(i(\omega_j - \omega)t) dt,$$

the conclusion follows from Lemma 5.53. \square

In particular, Theorem 5.52 yields the following.

5.54 Corollary. *If the finite sum of complex sinusoidal functions is zero in \mathbb{R} , then the coefficients of all components are zero.*

Another proof of Corollary 5.54. We give here a direct alternative proof based on induction on the number of components. The claim is trivial for $n = 1$. Let us prove it for $n = 2$. Suppose

$$c_1 \exp(i\omega_1 t) + c_2 \exp(i\omega_2 t) = 0 \quad \forall t, \quad \omega_1 \neq \omega_2.$$

Multiplying by $\exp(-i\omega_2 t)$ we find $c_1 \exp(i(\omega_1 - \omega_2)t) + c_2 = 0$ and differentiating

$$ic_1(\omega_1 - \omega_2) \exp(i(\omega_1 - \omega_2)t) = 0 \quad \forall t$$

which yields $c_1 = 0$ since $\omega_2 \neq \omega_1$. Consequently also $c_2 = 0$.

Suppose the theorem holds for the sum of $n - 1$ complex harmonic functions and let

$$f(t) = \sum_{j=1}^n c_j \exp(i\omega_j t) = 0.$$

Multiply by $\exp(-i\omega_n t)$ and differentiating we find

$$\sum_{j=1}^{n-1} ic_j(\omega_j - \omega_n) \exp(i\omega_j t) = 0;$$

therefore $c_j(\omega_j - \omega_n) = 0$ for all $j = 1, \dots, n - 1$, because of the inductive assumption. Since $\omega_j \neq \omega_n$, we infer $c_j = 0$ for all $j = 1, \dots, n - 1$ and consequently also $c_n = 0$. \square

Though sinusoidal signals are periodic, *finite sums of sinusoidal signals are periodic if and only if the sum is a trigonometric polynomial*. In fact, we have the following.

5.55 Theorem. *The sum of sinusoidal signals is periodic if and only if the quotients of the pulses of the components are rational.*

Proof. Let $f(t) = \sum_{j=1}^n c_j \exp(i\omega_j t)$ be the sum of sinusoidal signals with $\omega_i \neq \omega_j$ for $i \neq j$ and $c_j \neq 0$. Write $T_j = 2\pi/\omega_j$ for the period of the j -th component.

We may and do assume that $\omega_1 \neq 0$. If $\omega_j = \omega_1 p_j/q_j$ for $j = 1, \dots, n$, then $p_j T_j := q_j T_1$. Each T_j is then a submultiple of $T := q_2 \cdot q_3 \cdots q_n T_1$. Consequently every component is periodic of period T , hence $f(t)$ is periodic of period T .

Conversely suppose that $f(t)$ is periodic of period $T > 0$. We have

$$f(t+T) - f(t) = \sum_{j=1}^n c_j (\exp(i\omega_j T) - 1) \exp(i\omega_j t) = 0,$$

then Corollary 5.54 implies $\exp(i\omega_j T) = 1$. It follows that for every $j = 1, \dots, n$, there exists an integer k_j such that $\omega_j T = 2k_j\pi$, i.e., the component's pulses have rational quotients. \square

5.5 Summing Up

Polynomials

Let $A, B \in \mathbb{K}[x]$ be two polynomials with coefficients in the field \mathbb{K} . B divides A if $A = BQ$. A is said to be *irreducible* if no polynomial divides A but the constants. Two polynomials A and B are coprime if they do not have a common divisor but the constants. The greatest common divisor of A and B , g.c.d. (A, B) , is the divisor of A and B of highest degree. All greatest common divisors differ by a multiplicative constant, and one of them can be quickly computed by Euclid's algorithm.

- DIVISION ALGORITHM. Assume that A and B are two polynomials in $\mathbb{K}[x]$ with $\deg A < \deg B$; then there exist uniquely defined polynomials Q and R such that $A(x) = B(x)Q(x) + R(x)$ with $\deg R < \deg B$.
- BEZOUT IDENTITY. Given two polynomials A and B , there exist polynomials $U, V \in \mathbb{K}[x]$ such that $A(x)U(x) + B(x)V(x) = \text{g.c.d.}(A, B)(x)$.
- UNIQUE FACTORIZATION THEOREM. Any polynomial in $\mathbb{K}[x]$ is the product of its irreducible factors. The decomposition is unique apart from the order of the factors.
- FACTOR THEOREM. $\alpha \in \mathbb{K}$ is a root of $P \in \mathbb{K}$, i.e., $P(\alpha) = 0$, if and only if $x - \alpha$ divides $P(x)$. Therefore $P(x)$ has at most n distinct roots if $\deg P = n$.
- PRINCIPLE OF IDENTITY OF POLYNOMIALS. Two polynomials P, Q with degree at most n must be equal if their polynomial functions coincide on $n+1$ distinct points in \mathbb{K} .
- FUNDAMENTAL THEOREM OF ALGEBRA A polynomial $P(z) = \sum_{j=0}^n a_j z^j$ of degree n in $\mathbb{C}[z]$ factorizes as a product of n polynomials of first degree

$$P(z) = a_n(z - \alpha_1)(z - \alpha_2) \cdots (z - \alpha_n) \quad \text{in } \mathbb{C}[z].$$

- FACTORIZATION IN $\mathbb{R}[x]$. A polynomial with real coefficients of degree n in $\mathbb{R}[x]$ factorizes as a product of first and second order irreducible polynomials in $\mathbb{R}[x]$.

Polynomial equations

There exist algebraic procedures that fully solve in \mathbb{C} the polynomial equations of *third* and *fourth* degree, see 5.21 and 5.23. Polynomial equations of degree higher than five cannot be *solved by radicals* in general. There are however simpler rules to compute the number of real, positive roots of a real polynomial, see Theorem 5.25, and the number of zeros of a real polynomial in a given interval $[a, b]$ of \mathbb{R} , see Theorem 5.27.

Rational functions

Rational functions are quotients of complex polynomials

$$R(z) := \frac{C(z)}{D(z)}.$$

They are defined on $\mathbb{C} \setminus \{z \mid D(z) = 0\}$. By the division algorithm, $R(z)$ decomposes as a sum of a polynomial, which is nonzero if and only if $\deg C \geq \deg D$, and of a rational function $S(z) := A(z)/B(z)$ such that $\deg A < \deg B$. We can now decompose $A(z)/B(z)$ in simpler fractions, once the roots of B are known.

◦ HERMITE'S DECOMPOSITION FORMULA IN \mathbb{C} . We have

$$\frac{A(z)}{B(z)} = \sum_{\alpha \text{ root of } B} \sum_{j=1}^{k(\alpha)} \frac{\lambda_{\alpha,j}}{(z-\alpha)^j} \quad (5.39)$$

where the $\lambda_{\alpha,j}$ can be computed as follows: Set $Q_\alpha(z)$ be such that $B(z) = (z-\alpha)^{k(\alpha)}Q_\alpha(z)$. We then have $Q_\alpha(\alpha) \neq 0$. Set $R_{\alpha,k(\alpha)}(z) := A(z)$, then compute iteratively for $j = k(\alpha), \dots, 2, 1$,

$$\begin{cases} \lambda_{\alpha,j} := R_{\alpha,j}(\alpha)/Q_\alpha(\alpha), \\ R_{\alpha,j-1}(z) := (R_{\alpha,j} - \lambda_{\alpha,j}Q_\alpha(z))/(z-\alpha). \end{cases}$$

◦ HERMITE'S DECOMPOSITION FORMULA IN \mathbb{R} . Let $A(x)$ and $B(x) \in \mathbb{R}[x]$ with $\deg A < \deg B$. Let $\deg B =: n$ and denote by $k(\alpha)$ the multiplicity of the root α of B . Then

$$\frac{A(x)}{B(x)} = \sum_{\substack{\alpha \text{ root of } B \\ \alpha \in \mathbb{R}}} \sum_{j=1}^{k(\alpha)} \frac{\lambda_{\alpha,j}}{(x-\alpha)^j} + 2 \sum_{\substack{\alpha \text{ root of } B \\ \Im(\alpha) > 0}} \sum_{j=1}^{k(\alpha)} \Re\left(\frac{\lambda_{\alpha,j}}{(x-\alpha)^j}\right) \quad (5.40)$$

for all $x \in \mathbb{R}$ such that $B(x) \neq 0$. The constants $\lambda_{\alpha,j}$ are the same as in (5.39)

Trigonometric polynomials

A (complex) trigonometric polynomial is a function of the type

$$P(t) = \sum_{k=-n}^n c_k \exp\left(i \frac{2\pi k}{T} t\right), \quad t \in \mathbb{R}$$

where $T > 0$ and, for $k = -n, \dots, n$, $c_k \in \mathbb{C}$. The numbers c_k are called the *spectrum* of P and obviously fix P .

◦ The spectrum can be recovered from P by

$$c_k = \frac{1}{T} \int_0^T P(t) e^{-i \frac{2\pi k}{T} t} dt, \quad k = -n, \dots, n,$$

◦ the *energy equality* holds

$$\frac{1}{T} \int_0^T |P(t)|^2 dt = \sum_{k=-n}^n |c_k|^2.$$

◦ SAMPLING. The trigonometrical polynomial $P(t) \in \mathcal{P}_{n,2\pi}$ and its spectrum $\{c_k\}$, $k \in \{-n, \dots, n\}$, can be computed by *sampling* P at the points $t_k := 2\pi k/(2n+1)$, $k \in \{-n, \dots, n\}$. We have

$$P(t) = \frac{1}{2n+1} \sum_{j=-n}^n P(t_j) D_n(t-t_j) \quad \forall t \in \mathbb{R},$$

$$c_k = \frac{1}{2n+1} \sum_{j=-n}^n P(t_j) e^{ikt_j}, \quad k = -n, \dots, n$$

where

$$D_n(t) := \sum_{k=-n}^n e^{ikt} = \frac{\sin((n+1/2)t)}{\sin(t/2)}$$

is the *Dirichlet's kernel* of order n .

5.6 Exercises

5.56 ¶. Let $P(z) = a_2 z^2 + a_1 z + a_0$ and let α_1, α_2 be its two complex roots. Show that

$$a_1/a_2 = \alpha_1 + \alpha_2, \quad a_0/a_2 = \alpha_1 \alpha_2.$$

5.57 ¶. Let $P(z) = a_3 z^3 + a_2 z^2 + a_1 z + a_0$ and let $\alpha_1, \alpha_2, \alpha_3$ be its three complex roots. Show that

$$\alpha_1 + \alpha_2 + \alpha_3 = -a_2/a_3, \quad \alpha_1 \alpha_2 + \alpha_1 \alpha_3 + \alpha_2 \alpha_3 = a_1/a_3, \quad \alpha_1 \alpha_2 \alpha_3 = -a_0/a_3.$$

5.58 ¶. Suppose that the coefficients of $a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$ are integers and that $a_n = 1$. Suppose that such a polynomial has a real root x . Show that either x is an integer or x is irrational. In the case x is an integer, notice that x divides a_0 .

5.59 ¶. Suppose that the equation $x^3 + px^2 + qx + r = 0$ has three real roots. and let d be the difference between the largest and the smallest root. Show that

$$\sqrt{p^2 - 3q} \leq d \leq \frac{2}{\sqrt{3}} \sqrt{p^2 - 3q}.$$

5.60 ¶. Let x_0 be a root of $x^n + a_{n-1} x^{n-1} + \dots + a_0 = 0$. Show that $|x_0| \leq 1 + |a_0| + \dots + |a_{n-1}|$. [*Hint:* Consider separately the case $|x_0| < 1$ and $|x_0| \geq 1$.]

5.61 ¶. Let $P(z) = \sum_{k=0}^n a_k z^k$ be a polynomial of degree n and let $\alpha_1, \dots, \alpha_n$ be the n roots. Show that

$$\sigma_k(\alpha_1, \dots, \alpha_n) = (-1)^k a_{n-k}/a_n, \quad k = 1, \dots, n$$

where $\sigma_k(\alpha_1, \dots, \alpha_n)$ is the sum of all possible products of k roots, called the *symmetric k -function*.

5.62 ¶. Every real polynomial which is nonnegative for all real x may be written in the form $P^2(x) + Q^2(x)$ where P and Q are real polynomials. [*Hint:* Observe $(p^2 + q^2)(r^2 + s^2) = (pr + qs)^2 + (ps - qr)^2$.]

5.63 ¶ Lagrange's interpolating polynomials. Given $n+1$ points in \mathbb{C} , x_0, x_1, \dots, x_n and $n+1$ values y_0, y_1, \dots, y_n in \mathbb{C} , show that a polynomial \bar{P} of degree n such that

$$\bar{P}(x_i) = y_i \quad i = 0, 1, \dots, n$$

necessarily has the form

$$\bar{P}(x) = \sum_{j=0}^n L_j^n(x) y_j$$

where the $L_j^n(x)$ are the unique polynomials of degree n , called *Lagrange's interpolating polynomials*, such that

$$L_j^n(x_i) = \delta_{ij}, \quad i, j = 0, 1, \dots, n,$$

δ_{ij} being the Kronecker symbol. Write explicitly the $L_j^n(x)$'s.

5.64 ¶¶ Trigonometric solution of third degree equations. Consider the equation $x^3 + px + q = 0$ where $p, q \in \mathbb{R}$.

- (i) If $q^2/4 + p^3/27 < 0$ and $p > 0$, replace x with rx to obtain

$$y^3 + \frac{p}{r^2}y + \frac{q}{r^3} = 0.$$

Comparing with the trigonometric identity

$$\cos^3 \frac{\varphi}{3} - \frac{3}{4} \cos \frac{\varphi}{3} - \frac{1}{4} \cos \varphi = 0,$$

write the roots in terms of trigonometric functions of φ .

- (ii) If $q^2/4 + p^3/27 > 0$ and $p > 0$, set

$$\tan \varphi := -\left(\sqrt{\frac{p}{3}}\right)^3 \frac{2}{q}, \quad 0 < \varphi < \pi, \quad \tan \psi := \sqrt[3]{\tan \frac{\varphi}{2}},$$

and find the roots as functions of ψ .

- (iii) If $q^2 + p^3/27 > 0$ and $p < 0$, set

$$\sin \varphi := \left(\sqrt{-\frac{p}{3}}\right)^2 \left(-\frac{2}{q}\right), \quad 0 < \varphi < \pi$$

and find the roots in function of φ .

5.65 ¶. A hydrostatic approach to solve third order equations was proposed in 1898 by A. Demanet. Consider two communicating vessels, one being a circular cone of radius r and altitude a , the other a cylinder with basis of 1 square centimeter. If

$$\frac{r}{a} = \sqrt{\frac{3}{\pi}},$$

and if h is the altitude that is reached by c cubic centimeters of liquid, show that h solves the equation

$$x^3 + x = c.$$

6. Series

Processes of *infinite summation* or *infinite series*, or, simply, *series* have appeared since ancient times. Aristotle (384BC–322BC) in his *Physics* seems to be aware that the geometric series

$$1 + q + q^2 + q^3 + \cdots$$

has a finite sum if $|q| < 1$. Later François Viète (1540–1603) in fact computed (in 1593)

$$1 + q + q^2 + q^3 + \cdots = \frac{1}{1 - q}.$$

Zeno's paradox of dichotomy clearly concerns the decomposition of 1 into the infinite series

$$1 = \frac{1}{2} + \frac{1}{2^2} + \frac{1}{2^3} + \cdots.$$

In medieval times, Nicole d' Oresme (1323–1382) showed that the *harmonic series*

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

diverged. However, it was in the seventeenth century with the Calculus of Sir Isaac Newton (1643–1727) and Gottfried von Leibniz (1646–1716) that infinite series pervaded mathematics tremendously, especially as *power series*.

For Gottfried von Leibniz (1646–1716) and Sir Isaac Newton (1643–1727) and their contemporaries such as John Wallis (1616–1703), James Gregory (1638–1675), Brook Taylor (1685–1731), James Stirling (1692–1770), and Colin MacLaurin (1698–1746), functions were essentially *infinite polynomials* or *power series*, and with them one operated to differentiate and to integrate or compute areas, to calculate special quantities such as e and π and the logarithmic and trigonometric functions, to interpolate series of data (which was particularly useful for navigation). For example, Leibniz found

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots,$$

the right-hand side of which is now called a *Leibniz series*, and Newton, in *De analysi per equationes numero terminorum infinitas*, found the series of $\log(1 + x)$, $\sin x$, $\cos x$, $\arcsin x$, e^x , \dots as, for instance,

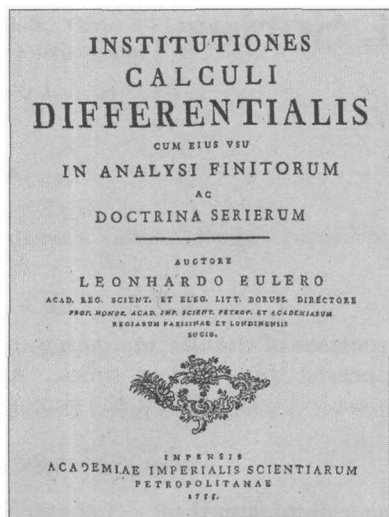
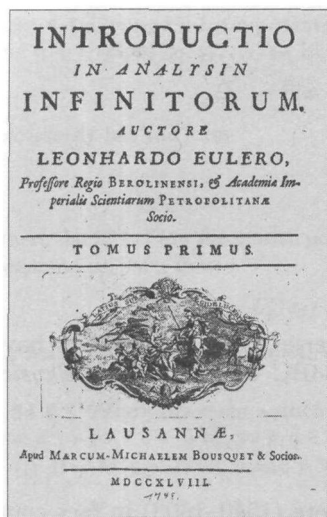


Figure 6.1. Frontispieces of *Introductio* ... and of *Institutiones calculi* ... by Leonhard Euler (1707–1783).

$$\log(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 + \cdots,$$

from which

$$\log 2 = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots;$$

James Gregory (1638–1675) found

$$\arctan x = x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots$$

nowadays called a *Gregory series*, the special case of which with $x = 1$ is a Leibniz series.

In this chapter we illustrate methods for the study of the convergence of numerical series, while in the next we shall deal with basic facts concerning *power series*.

6.1 Basic Facts

Given a sequence $\{a_n\}$, $n = 0, 1, 2, \dots$, of real or complex numbers, the recurrence

$$\begin{cases} s_0 = a_0, \\ s_{n+1} = s_n + a_{n+1}, \quad \forall n \geq 0 \end{cases} \quad (6.1)$$

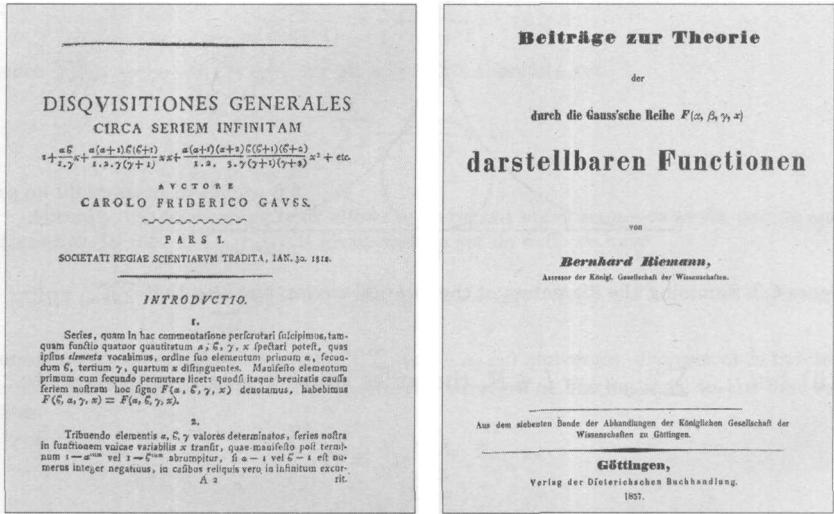


Figure 6.2. The first page of the paper by Carl Friedrich Gauss (1777–1855) and the frontispiece of G. F. Bernhard Riemann's (1826–1866) paper on hypergeometric series.

defines by induction a *unique sequence* $\{s_n\}$,

$$s_n := a_0 + a_1 + a_2 + \cdots + a_n = \sum_{j=0}^n a_j,$$

see Example 2.5. The sequence $\{s_n\}$ is called the *sequence of partial sums* of $\{a_n\}$ or just the *series* of $\{a_n\}$. We use the symbol

$$\sum_{j=0}^{\infty} a_j$$

in order to indicate the sequence $\{s_n\}$. The presumption *let us consider the series* $\sum_{j=0}^{\infty} a_j$, is therefore a shorthand for *let us consider the sequence of partial sums* $\{\sum_{j=0}^n a_j\}$ of the sequence $\{a_n\}$.

a. Definitions and examples

Let $\sum_{n=0}^{\infty} a_n$, $a_n \in \mathbb{R}$, be a series of real numbers, and let $\{s_n\}$, $s_n := \sum_{k=0}^n a_k$, be the sequence of partial sums.

6.1 Definition. If $\{s_n\}$ converges to L , we say that the series converges and that L is the sum of the series.

Actually, three alternative situations are possible:

- (i) $\lim_{n \rightarrow \infty} \sum_{j=0}^n a_j$ does not exist; we then say that the *series is indeterminate*. This is the case of the series $\sum_{n=0}^{\infty} (-1)^n$.

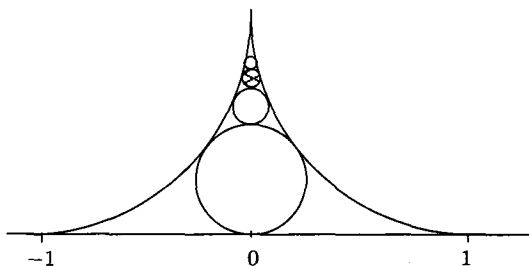


Figure 6.3. Summing the diameters of the internal circles, one sees that $\sum_{k=1}^{\infty} \frac{1}{k(k+1)} = 1$.

(ii) $\lim_{n \rightarrow \infty} \sum_{j=0}^n a_j = L \in \mathbb{R}$, the series *converges to L* and we write

$$\sum_{n=0}^{\infty} a_n = L$$

for $\sum_{j=0}^n a_j \rightarrow L$, $n \rightarrow \infty$.

(iii) $\lim_{n \rightarrow \infty} \sum_{j=0}^n a_j = +\infty$ (resp. $-\infty$). In this case we say that the *series diverges to $+\infty$ (resp. $-\infty$)* and we write

$$\sum_{n=0}^{\infty} a_n = +\infty \quad (\text{resp. } \sum_{n=0}^{\infty} a_n = -\infty).$$

6.2 ¶. Show that $\sum_{j=1}^{\infty} j$ and $\sum_{j=1}^{\infty} j^2$ diverge to $+\infty$.

6.3 Example (Geometric series). We saw in Example 2.65 that

$$G_q(n) := \sum_{j=0}^n q^j = \begin{cases} n+1 & \text{if } q = 1, \\ \frac{q^{n+1} - 1}{q - 1} & \text{otherwise,} \end{cases}$$

consequently the geometric series

$$\sum_{j=0}^{\infty} q^j \quad \begin{cases} \text{converges to } \frac{1}{1-q} & \text{if } |q| < 1, \\ \text{diverges to } +\infty & \text{if } q \geq 1, \\ \text{is indeterminate} & \text{if } q \leq -1. \end{cases}$$

6.4 Example (Telescoping series). These are series for which the general term a_n can be expressed in the form

$$a_n = b_n - b_{n-1}$$

for a suitable sequence $\{b_n\}$. In this case

$$\sum_{j=1}^n a_j = (b_n - b_{n-1}) + (b_{n-1} - b_{n-2}) + \cdots + (b_3 - b_2) + (b_2 - b_1) = b_n - b_1.$$

An example is given by *Mengoli's series* $\sum_{j=1}^{\infty} \frac{1}{j(j+1)}$, named after Pietro Mengoli (1626–1686). In fact

$$\frac{1}{j(j+1)} = \frac{1}{j} - \frac{1}{j+1}, \quad \forall j \geq 1,$$

hence $\sum_{j=1}^n \frac{1}{j(j+1)} = 1 - \frac{1}{n+1}$ for all $n \geq 1$. We therefore get

$$\sum_{j=1}^{\infty} \frac{1}{j(j+1)} = 1;$$

see an illustration in Figure 6.3.

Actually, the telescoping trick allows us to regard *every* sequence as the partial sum of a series. In fact, if $\{a_n\}_{n \geq 1}$ is given and we set $a_0 = 0$, we have

$$\sum_{j=1}^n (a_j - a_{j-1}) = a_n - a_0 = a_n, \quad \forall n \geq 1,$$

consequently we see that the series $\sum_{j=0}^{\infty} (a_j - a_{j-1})$ converges, diverges or is indeterminate according to whether $\{a_n\}$ converges, diverges or has no limit. In the first two cases

$$\sum_{j=1}^{\infty} (a_j - a_{j-1}) = \lim_{n \rightarrow \infty} a_n.$$

6.5 Example (Arithmetic-geometric series). There is a closed formula also for

$$S(n) := \sum_{j=1}^n j q^j, \quad n \geq 1, \quad q \in \mathbb{C}, \quad q \neq 1.$$

In fact, multiplying $S(n)$ by $1 - q^2$, we get

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

that yields

$$\sum_{j=1}^n j q^j = \frac{n q^{n+2} - (n + 1) q^{n+1} + q}{(1 - q)^2}.$$

Since $n q^n \rightarrow 0$ as $n \rightarrow \infty$ if and only if $|q| < 1$, see Example 2.59, $\sum_{j=0}^{\infty} j q^j$ converges if and only if $|q| < 1$ and in this case

$$\sum_{j=0}^{\infty} j q^j = \frac{q}{(1 - q)^2}.$$

6.6 Infinite product. We can define the *infinite product* of a sequence of positive numbers $\{a_n\}$,

$$\prod_{i=1}^{\infty} a_i := \lim_{n \rightarrow \infty} \prod_{i=1}^n a_i,$$

when it exists. Trivially, $\prod_{i=1}^{\infty} a_i$ exists if and only if the series $\sum_{i=1}^{\infty} \log a_i$ converges and

$$\prod_{i=1}^{\infty} a_i = \exp \left(\sum_{i=1}^{\infty} \log a_i \right).$$

b. A necessary condition for convergence

6.7 Proposition. If $\sum_{n=0}^{\infty} a_n$ converges, then $a_n \rightarrow 0$.

Proof. In fact

$$a_n = \sum_{j=0}^n a_j - \sum_{j=0}^{n-1} a_j \rightarrow L - L = 0,$$

if L is the sum of $\sum_{n=0}^{\infty} a_n$. □

However, the condition $a_n \rightarrow 0$ is not sufficient to ensure convergence of $\sum_{n=0}^{\infty} a_n$. For instance, we shall see in Example 6.27 that the *harmonic series* $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges.

c. Series and improper integrals

The concept of sum of a series can be seen as a particular case of an improper integral (see Section 4.5.2 of [GM1]), and this is quite a useful remark, see, for instance Example 6.25. To a sequence $\{a_n\}$, $n \geq 0$, of real numbers, we associate the piecewise constant function $\varphi : [0, +\infty[\rightarrow \mathbb{R}$ defined by

$$\varphi_a(x) = a_n \quad \text{if } n \leq x < n+1.$$

Clearly φ is measurable and, for all $n \geq 0$ we have

$$\sum_{j=0}^n a_j = \int_0^{n+1} \varphi_a(x) dx. \quad (6.2)$$

Actually

6.8 Proposition. The sequence of partial sums of $\{a_n\}$ has a limit in $\overline{\mathbb{R}}$ if and only if $\int_0^x \varphi_a(x) dx$ has a limit when $x \rightarrow +\infty$. In this case

$$\sum_{j=0}^{\infty} a_j = \lim_{x \rightarrow \infty} \int_0^x \varphi_a(x) dx.$$

In particular $\sum_{j=0}^{\infty} a_j$ converges if and only if φ has an improper integral at infinity.

6.9 ¶. Prove Proposition 6.8. [*Hint:* Show that $\lim_{x \rightarrow +\infty} \int_0^x \varphi_a(x) dx$ exists if and only if $\lim_{n \rightarrow \infty} \int_0^n \varphi_a(x) dx$ exists and

$$\lim_{x \rightarrow +\infty} \int_0^x \varphi_a(x) dx = \lim_{n \rightarrow \infty} \int_0^n \varphi_a(x) dx.$$

For that, set $\phi(x) := \int_0^x \varphi(x) dx$ and observe that for $n := [x]$,

$$\begin{cases} \phi(n) \leq \phi(x) \leq \phi(n+1) & \text{if } a_n \geq 0, \\ \phi(n+1) \leq \phi(x) \leq \phi(n) & \text{if } a_n \leq 0. \end{cases} \quad (6.3)$$

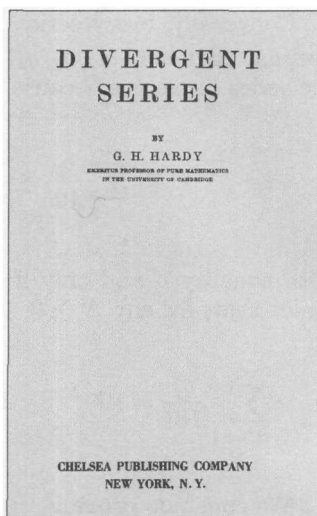
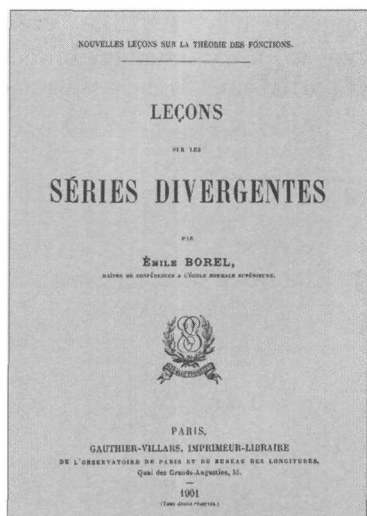


Figure 6.4. Frontispieces of two books respectively by Emile Borel (1871–1956) and G. H. Hardy (1877–1947) on *divergent series*.

d. Decimals

Every real number has a *decimal representation* $x = q_0, q_1, q_2, q_3, \dots$, which is defined iteratively by

$$x_0 := x, \quad q_0 := [x_0], \quad \begin{cases} x_{j+1} := x_j - \frac{q_j}{10^j} \\ q_{j+1} := [10^{j+1}x_{j+1}] \end{cases} \quad \forall j \geq 0$$

where $[\alpha]$ is the largest integer not greater than α . Inductively we also see that $q_j \in \{0, \dots, 9\}$ for $j > 0$ and that $0 \leq x_j < 10^{-j+1}$. In particular $x_j \rightarrow 0$ as $j \rightarrow \infty$. Moreover

$$q_0 + \sum_{j=1}^N \frac{q_j}{10^j} = q_0 + \sum_{j=1}^N (x_j - x_{j+1}) = q_0 + x_1 - x_{N+1} = x - x_{N+1}$$

hence, when $N \rightarrow \infty$

$$x = \sum_{j=0}^{\infty} \frac{q_j}{10^j} = q_0 + \frac{q_1}{10} + \frac{q_2}{100} + \frac{q_3}{1000} + \dots,$$

or, as we commonly write, $x = q_0, q_1, q_2, q_3, \dots$.

The algorithm giving the decimal alignment may stop after N steps, i.e., $x_j = q_j = 0 \quad \forall j > N$. This clearly happens if and only if $x = \frac{k}{10^N}$, $k \in \{0, \dots, 10^N - 1\}$. However, even for rational numbers the decimal alignment may be infinite as for $1/3 = 0, 333333 \dots$.

Conversely, every series $\sum_{j=0}^{\infty} q_j/10^j$, $q_j \in \{0, 1, 2, \dots, 9\}$, i.e., every decimal alignment q_0, q_1, q_2, \dots , is (converges to) a real number. In fact, the series converges because the sequence of partial sums is increasing and

$$\sum_{j=1}^{\infty} \frac{q_j}{10^j} \leq \sum_{j=1}^{\infty} \frac{9}{10^j} = \frac{9}{10} \frac{1}{1 - 1/10} = 1$$

with equality if and only if $q_j = 9 \ \forall j \geq 1$. The same reasoning actually yields that, for any $N \geq 0$,

$$\sum_{j=N+1}^{\infty} \frac{q_j}{10^j} = 10^{-N} \quad \text{if and only if} \quad q_j = 9 \ \forall j \geq N+1. \quad (6.4)$$

We conclude proving

6.10 Proposition. *Two decimal alignments q_0, q_1, q_2, \dots and h_0, h_1, h_2, \dots have different sums (represent different numbers) except when there exists an integer N such that*

$$\text{either} \quad \begin{cases} q_N = 1 + h_N, \\ q_j = 0, \ h_j = 9 \ \forall j > N, \end{cases} \quad \text{or} \quad \begin{cases} h_N = q_N + 1, \\ q_j = 9, \ h_j = 0 \ \forall j > N. \end{cases}$$

Proof. Suppose that $q_0 + \sum_{j=1}^{\infty} \frac{q_j}{10^j} = h_0 + \sum_{j=1}^{\infty} \frac{h_j}{10^j}$, i.e.,

$$q_0 - h_0 + \sum_{j=1}^{\infty} \frac{q_j - h_j}{10^j} = 0.$$

If $q_j = h_j$ does not hold $\forall j$, denote by N the first index for which $q_N \neq h_N$. Then

$$\frac{q_N - h_N}{10^N} = - \sum_{j=N+1}^{\infty} \frac{q_j - h_j}{10^j} =: -R_N$$

and $|q_N - h_N| = 1$ since $|R_N| \leq 10^{-N}$. If for instance $h_N = 1 + q_N$, then

$$R_N = \sum_{j=N+1}^{\infty} \frac{q_j - h_j}{10^j} = 1;$$

(6.4) then yields $q_j - h_j = 9$, that is, $q_j = 9$ and $h_j = 0$ for all $j \geq 1$. □

6.11 Example. 0.234765799999... and 0.2347658 are the same rational number.

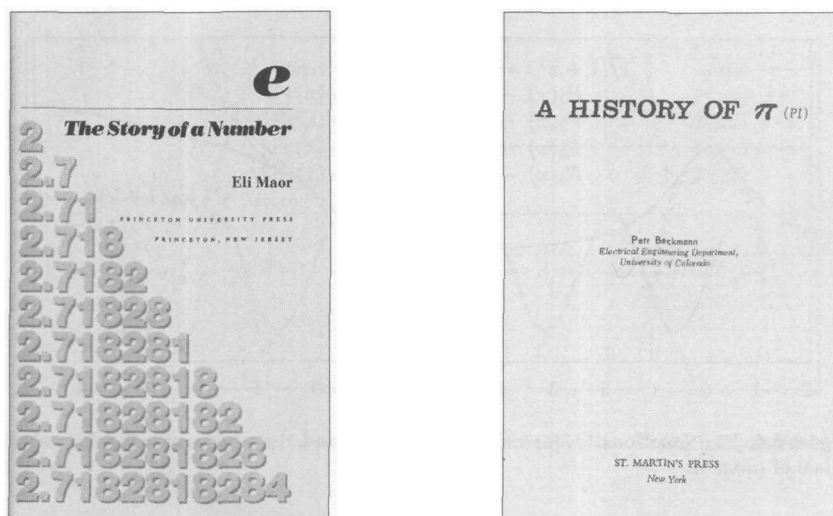


Figure 6.5. Two popular books on e and π .

6.2 Taylor Series, e and π

A simple and natural way of generating convergent power series is by starting with a function $f :] - a, a[\rightarrow \mathbb{R}$ of class C^∞ and considering its *Taylor's series*

$$\sum_{j=0}^{\infty} \frac{D^j f(0)}{j!} x^j, \quad x \in \mathbb{R},$$

which has as partial sums Taylor's polynomials

$$P_n(x) := \sum_{j=0}^n \frac{D^j f(0)}{j!} x^j.$$

Obviously

$$\sum_{j=0}^{\infty} \frac{D^j f(0)}{j!} x^j = f(x)$$

if and only if the remainder $R_n(x) := f(x) - P_n(x)$ tends to zero as $n \rightarrow \infty$. This happens to be true for quite a number of elementary functions, as we shall see in the following examples. However, in general

$$f(x) \neq \sum_{n=0}^{\infty} \frac{D^n f(0)}{n!} x^n \quad \forall x \neq 0,$$

as shown for instance by the following

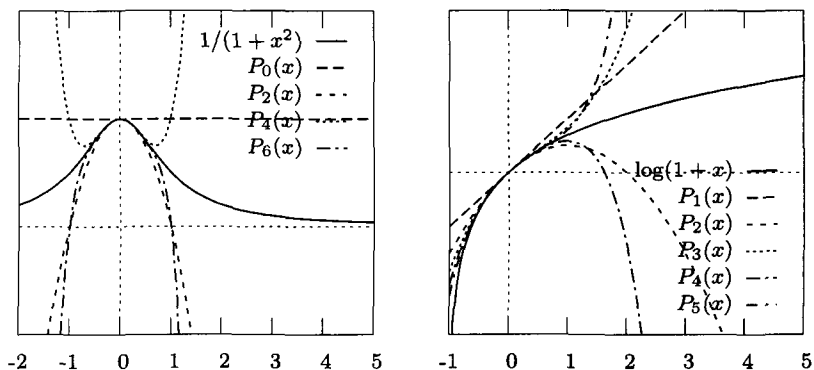


Figure 6.6. The functions $1/(1+x^2)$ and $\log(1+x)$ and their respective Taylor polynomials of order n .

6.12 Example. It is easily seen by induction that the function

$$f(x) = \begin{cases} \exp(-1/x^2) & \text{if } x \neq 0, \\ 0 & \text{if } x = 0 \end{cases}$$

has derivatives of any order and $D^j f(0) = 0 \forall j$. Its Taylor series sums to zero while $f(x) \neq 0$ for all $x \neq 0$.

Some of the following examples have already been discussed, see e.g, Section 5.1 of [GM1], but, for the reader's convenience, we repeat them here.

6.13 Example (Logarithm). Replacing x by $-x$ in the well-known identity

$$\frac{1}{1-x} = \sum_{k=0}^n x^k + \frac{x^{n+1}}{1-x}, \quad (6.5)$$

we get

$$\frac{1}{1+x} = \sum_{k=0}^n (-1)^k x^k + \frac{(-x)^{n+1}}{1+x}, \quad x \neq -1,$$

and integrating between 0 and x ,

$$\begin{aligned} \log(1+x) &= \int_0^x \frac{1}{1+t} dt = \int_0^x \sum_{k=0}^n (-1)^k t^k dt + R_n(x) \\ &= \sum_{k=0}^n (-1)^k \frac{x^{k+1}}{k+1} + R_n(x) \end{aligned}$$

where

$$R_n(x) := \int_0^x \frac{(-t)^{n+1}}{1+t} dt.$$

The remainder can be easily estimated for $x > -1$ by

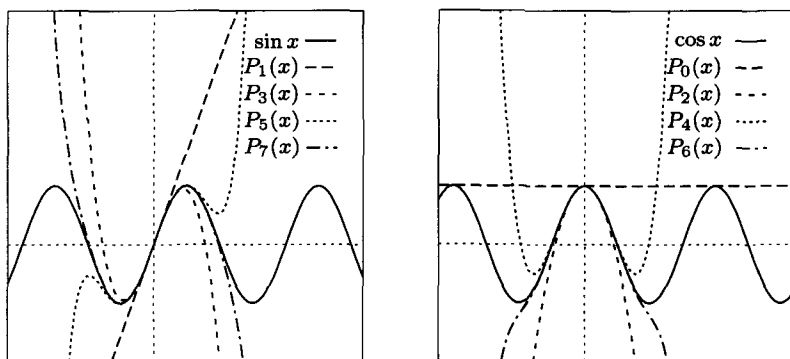


Figure 6.7. The functions $\sin x$ and $\cos x$ and their respective Taylor polynomials of order n .

$$|R_n(x)| \leq \max \left(1, \frac{1}{1+x} \right) \frac{|x|^{n+1}}{n+1}, \quad (6.6)$$

hence it converges to zero if $-1 < x \leq 1$. We then conclude that the Taylor series of $\log(1+x)$ converges if $-1 < x \leq 1$ and

$$\log(1+x) = \sum_{k=0}^{\infty} (-1)^k \frac{x^{k+1}}{k+1}, \quad -1 < x \leq 1. \quad (6.7)$$

6.14 Example (The arc tangent function). Starting from (6.5), replacing x by $-x^2$, and integrating we get

$$\begin{aligned} \arctan x &= \int_0^x \frac{1}{1+t^2} dt = \int_0^x \sum_{k=0}^{\nu} (-1)^k t^{2k} dt + R_n(x) \\ &= \sum_{k=0}^n (-1)^k \frac{x^{2k+1}}{2k+1} + R_n(x) \end{aligned}$$

where

$$R_n(x) := \int_0^x \frac{(-t^2)^{n+1}}{1+t^2} dt.$$

Since

$$|R_n(x)| \leq \left| \int_0^x t^{2n+2} dt \right| = \frac{|x|^{2n+3}}{2n+3}, \quad (6.8)$$

we infer that $R_n(x) \rightarrow 0$ as $n \rightarrow \infty$ if $|x| \leq 1$, concluding that

$$\arctan x = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1}, \quad |x| \leq 1. \quad (6.9)$$

6.15 Example (Taylor series for e^x , $\sin x$, $\cos x$). We know that Taylor's polynomial of degree n of e^x centered at 0 is

$$P_n(x) := \sum_{k=0}^n \frac{f^{(k)}(0)}{k!} x^k = \sum_{k=0}^n \frac{x^k}{k!}$$

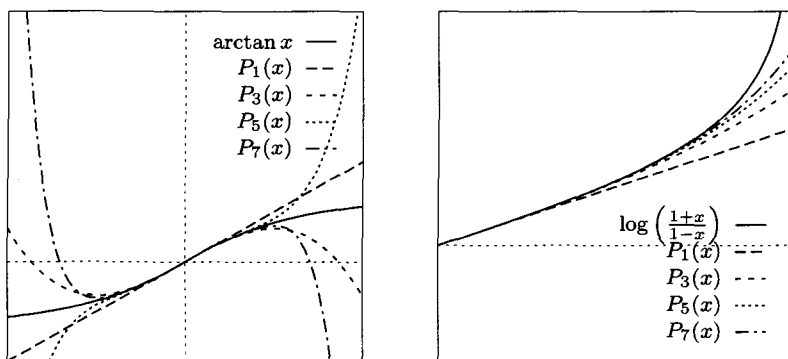


Figure 6.8. The functions $\arctan x$ and $\log\left(\frac{1+x}{1-x}\right)$ and their respective Taylor polynomials of order n .

and Taylor's formula with Lagrange remainder, see e.g., 5.5 of [GM1], yields

$$e^x - \sum_{k=0}^n \frac{x^k}{k!} = \frac{e^{\xi_n}}{(n+1)!} x^{n+1}$$

for a suitable ξ_n in the interval $]0, x[$ (or $]x, 0[$ is $x < 0$). Therefore, for any $x \in \mathbb{R}$,

$$\left| e^x - \sum_{k=0}^n \frac{x^k}{k!} \right| \leq \max(e^x, 1) \frac{|x|^{n+1}}{(n+1)!} \rightarrow 0 \quad \text{as } n \rightarrow \infty, \quad (6.10)$$

thus concluding that the series $\sum_{n=0}^{\infty} \frac{x^n}{n!}$ converges and

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \quad \forall x \in \mathbb{R}. \quad (6.11)$$

Similarly, see Figure 6.7, one proves that

$$\sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} = \sin x, \quad \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} = \cos x, \quad \forall x \in \mathbb{R}. \quad (6.12)$$

6.16 ¶. Prove (6.12).

a. The number π

Approximated values of π have been known since ancient times. The first twenty digits are

$$\pi = 3.14159265358979323846 \dots$$

The first analytic representation of π was probably found by François Viète (1540–1603) in 1579 as the infinite product

Book of kings	$\pi \sim 3$
Arithmetic book by Ahmes (1900 a. C.)	$\pi \sim (16/9)^2 \sim 3.16$
Śālbāṣūtras (500 a. C.)	$\pi \sim (26/15)^2 \sim 3.0044$
Plato (V sec. a. C.)	$\pi \sim \sqrt{2} + \sqrt{3} \sim 3.14626$
Archimedes (III sec. a. C.) provides estimates from above and from below by means of inscribed and circumscribed polygons of 96 sides	$3 + \frac{10}{71} < \pi < 3 + \frac{1}{7}$
Zhang Heng (I sec. . C.)	$\pi \sim \sqrt{10} \sim 3.162$
Ptolemy's (~ 150 d. C.) takes in the Archimedes approximation	$\pi \sim 3 + \frac{17}{120} \sim 3.14166$
Wang Fang (II sec. d. C.)	$\pi \sim \frac{142}{45} \sim 3.155$
Liu Hui (~ 263 d. C.) estimates with polygons of 192 sides	$3.14 + \frac{64}{62500} < \pi < 3.14 + \frac{169}{62500}$
Liu Hui (~ 263 d. C.) estimates with polygons of 3072 sides	$\pi \sim 3.14159$
Zhu Chong-Zhi (430–501) finds π up to six digits with a convergent in the continuous fraction development	$\pi \sim \frac{355}{113} \sim 3.1415929$
Ārayabhata (498 d. C.) and al-Hwārizmī (IX sec. d. C.)	$\pi \sim \frac{62832}{20000} = 3.1416$
Leonardo Pisano (1170–1250), called Fibonacci estimates with polygons of 96 sides	$\pi \sim \frac{864}{275} \sim 3.141818$
Albrecht Dürer (1471–1528)	$\pi \sim 3 + \frac{1}{8}$
Ludolph Van Ceulen (1540–1610) finds π up to 35 digits	3.14159...

Figure 6.9. The values of π computed or estimated before the infinitesimal calculus.

$$\frac{2}{\pi} = \sqrt{\frac{1}{2}} \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}} \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}}} \cdots,$$

see (6.26); and John Wallis (1616–1703) in 1655 found

$$\frac{\pi}{2} = \frac{2 \cdot 2}{1 \cdot 3} \frac{4 \cdot 4}{3 \cdot 5} \frac{6 \cdot 6}{5 \cdot 7} \cdots \frac{2n \cdot 2n}{(2n-1)(2n+1)} \cdots;$$

see (6.29) below. In 1671 James Gregory (1638–1675) found the representation

$$\frac{\pi}{4} = 1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \cdots,$$

independently found also by Gottfried von Leibniz (1646–1716) in 1674. Computing the Taylor series of $\arcsin x$,

n	S_n	R_n	n	S_n	R_n
1	2.666666666666667	$5E-01$	1	3.079201435678004	$6E-02$
3	2.895238095238096	$2E-01$	2	3.156181471569954	$-1E-02$
10	3.232315809405594	$-9E-02$	3	3.137852891595680	$4E-03$
30	3.173842337190750	$-3E-02$	5	3.141308785462883	$3E-04$
100	3.151493401070991	$-1E-02$	7	3.141568715941784	$2E-05$
300	3.144914903558853	$-3E-03$	9	3.141590510938080	$2E-06$
1000	3.142591654339544	$-1E-03$	11	3.141592454287646	$2E-07$
3000	3.141925875839790	$-3E-04$	13	3.141592634547314	$2E-08$
10000	3.141692643590535	$-1E-04$	15	3.141592651733998	$2E-09$
π	3.141592653589793		π	3.141592653589793	

Figure 6.10. The partial sums $S_n := \sum_{j=0}^n a_j$ and the error $R_n := \pi - S_n$: (a) on the left, for $a_j := 4(-1)^j/(2j+1)$; (b) on the right for $a_j := (-1)^j 2\sqrt{3} \frac{1}{3j(2j+1)}$.

$$\arcsin x = \sum_{n=0}^{\infty} \frac{1 \cdot 3 \cdots (2n-1)}{2 \cdot 4 \cdots 2n} \frac{x^{2n+1}}{2n+1}, \quad |x| \leq 1,$$

for $x = 1/2$. In 1665 Sir Isaac Newton (1643–1727) found

$$\frac{\pi}{6} = \frac{1}{2} + \frac{1}{2} \frac{1}{3} \frac{1}{8} + \frac{1}{2} \frac{3}{4} \frac{1}{5} \frac{1}{32} + \frac{1}{2} \frac{3}{4} \frac{5}{6} \frac{1}{7} \frac{1}{128} + \cdots.$$

The number π is the area of the unit circle that is, according to calculus,

$$\frac{\pi}{2} = \int_{-1}^1 \sqrt{1-x^2} dx$$

(see [GM1]), or the length of the halfcircle, that, as one can prove, is given by

$$\pi = \int_{-1}^1 \sqrt{1+y'(x)^2} dx = \int_{-1}^1 \frac{1}{\sqrt{1-x^2}} dx.$$

6.17 A few series that sum to π . From (6.9) and (6.8) we find the Leibniz–Gregory result

$$\frac{\pi}{4} = \arctan 1 = \sum_{n=0}^{\infty} (-1)^n \frac{1}{2n+1} \quad (6.13)$$

with the estimate for the rate of decay of the remainder

$$\left| \sum_{k=n+1}^{\infty} (-1)^k \frac{1}{2k+1} \right| = |R_n(1)| \leq \frac{1}{2n+3}.$$

However, the estimated rate of convergence is not fast, in accordance with the computed values in Figure 6.10 (a).

n	S_n	R_n
1	3.140597029326061	1E-03
2	3.141621029325035	-3E-05
3	3.141591772182178	9E-07
5	3.141592652615309	1E-09
7	3.141592653588603	1E-12
9	3.141592653589793	4E-16
π	3.141592653589793	

Figure 6.11. The partial sums $S_n := \sum_{j=0}^n a_j$ and the error $R_n := \pi - S_n$ for $a_j = 4(-1)^j \frac{1}{2j+1} \left(\frac{4}{5j+1} - \frac{1}{(239)j+1} \right)$.

A better approximation of π than (6.13) can be obtained in several ways. For instance, observing that $\frac{\pi}{6} = \arctan \frac{1}{\sqrt{3}}$, we get again from (6.9)

$$\frac{\pi}{6} = \sum_{n=0}^{\infty} (-1)^n \frac{1}{\sqrt{3}^{2n+1} (2n+1)},$$

i.e.,

$$\frac{\pi}{2\sqrt{3}} = \sum_{n=0}^{\infty} (-1)^n \frac{1}{3^n (2n+1)}. \quad (6.14)$$

If $\sqrt{3}$ is known, then (6.14) is a far better representation than (6.13), since (6.8) yields an exponential decay for the error,

$$\left| \sum_{k=n+1}^{\infty} (-1)^k \frac{1}{\sqrt{3}^{2k+1} (2k+1)} \right| = \left| R_n \left(\frac{1}{\sqrt{3}} \right) \right| \leq \frac{1}{\sqrt{3} (2n+3) 3^n},$$

see Figure 6.10 (b).

An even better approximation can be obtained using a simple trick. Recall that

$$\tan(\alpha + \beta) = \frac{\tan \alpha + \tan \beta}{1 + \tan \alpha \tan \beta}.$$

Starting with $\alpha := \arctan(1/5)$, we then get

$$\tan 2\alpha = \frac{5}{12}, \quad \tan 4\alpha = 1 + \frac{1}{119}, \quad \tan(4\alpha - \pi/4) = \frac{1}{239},$$

if we take into account that $4\alpha > \pi/4$. Hence

$$\frac{\pi}{4} = 4\alpha - (4\alpha - \pi/4) = 4 \arctan \frac{1}{5} - \arctan \frac{1}{239}$$

and conclude by (6.9) and (6.8)

$$\frac{\pi}{4} = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} \left(\frac{4}{5^{n+1}} - \frac{1}{(239)^{n+1}} \right) \quad (6.15)$$

with an exponential decay estimate for the remainder, see Figure 6.11.

n	S_n	R_n	n	S_n	R_n
1	1.0000000000000000	$-3E-01$	1	0.691358024691358	$2E-03$
3	0.8333333333333333	$-1E-01$	2	0.693004115226337	$1E-04$
10	0.645634920634921	$5E-02$	3	0.693134757332288	$1E-05$
30	0.676758137691398	$2E-02$	5	0.693147073759785	$1E-07$
100	0.688172179310195	$5E-03$	7	0.693147179548241	$1E-09$
300	0.691483291655625	$2E-03$	9	0.693147180549812	$1E-11$
1000	0.692647430559822	$5E-04$	11	0.693147180559840	$1E-13$
3000	0.692980541671060	$2E-04$	13	0.693147180559944	$1E-15$
10000	0.693097183059958	$5E-05$	15	0.693147180559945	$2E-16$
$\log 2$	0.693147180559945		$\log 2$	0.693147180559945	

Figure 6.12. The partial sums $S_n := \sum_{j=0}^n a_j$ and the errors $R_n := \log 2 - S_n$ for, on the left $a_j := (-1)^j/(j+1)$, and on the right $a_j := 2/(3^{2j+1}(2j+1))$.

6.18 Approximations of $\log 2$. Similarly, one can find a series which converges to $\log 2$. In fact, the Taylor series for $\log(1+x)$, (6.7) and (6.6), yield in particular

$$\log 2 = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots \quad (6.16)$$

and the estimate of the rate of decay for the remainder

$$\left| \log 2 - \sum_{k=1}^n \frac{(-1)^{k+1}}{k} \right| = |R_n(1)| \leq \frac{1}{n+1}.$$

This suggests that $R_n(1)$ decays slowly to zero, as in fact is the case, see Figure 6.12 (a).

A better approximation of $\log 2$ is obtained by observing that, for $0 < x < 1$,

$$\begin{aligned} \log \left(\frac{1+x}{1-x} \right) &= \log(1+x) - \log(1-x) \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1} - \sum_{n=0}^{\infty} (-1)^n \frac{(-x)^{n+1}}{n+1} \\ &= \sum_{n=0}^{\infty} \frac{x^{n+1}}{n+1} ((-1)^n + 1) = 2 \sum_{p=0}^{\infty} \frac{x^{2p+1}}{2p+1}, \end{aligned}$$

hence

$$\log 2 = \log \frac{1+1/3}{1-1/3} = 2 \sum_{n=0}^{\infty} \frac{1}{(2n+1)3^{2n+1}};$$

in this case the error decays exponentially, see Figure 6.12 (b), as

$$\sum_{k=n}^{\infty} \frac{1}{(2k+1)3^{2k+1}} \leq \frac{1}{3(2n+1)} \sum_{k=n}^{\infty} \frac{1}{9^k} = \frac{3}{8(2n+1)} \frac{1}{9^n}.$$

b. More on the number e

From (6.11) and (6.10) we infer

n	S_n	R_n
1	1.0000000000000000	$2E+00$
3	2.5000000000000000	$2E-01$
5	2.7083333333333333	$1E-02$
7	2.7180555555555555	$2E-04$
9	2.718278769841270	$3E-06$
11	2.718281801146385	$3E-08$
13	2.718281828286169	$2E-10$
15	2.718281828458230	$8E-13$
17	2.718281828459043	$2E-15$
e	2.718281828459045	

Figure 6.13. The partial sums $S_n := \sum_{j=0}^n 1/j!$ and the error $e - S_n$.

$$e = \sum_{n=0}^{\infty} \frac{1}{n!}, \quad (6.17)$$

with

$$\left| \sum_{k=n+1}^{\infty} \frac{1}{k!} \right| \leq \frac{e}{(n+1)!} < \frac{4}{(n+1)!}, \quad (6.18)$$

since $2 < e < 4$. For instance, for $n = 6$, $\sum_{j=0}^6 \frac{1}{j!} = 2.718055\dots$ approximates e from below with an error not higher than $4/7! = 1/1260$, which yields

$$2.718055\dots < e < 2.718849\dots$$

The estimate (6.18) also implies the following.

6.19 Theorem. e is irrational.

Proof. In fact, suppose on the contrary e rational, $e = p/q$, $p, q \in \mathbb{Z}$, $q \neq 0$. From (6.18)

$$\frac{p}{q} - \sum_{j=0}^n \frac{1}{j!} < \frac{4}{(n+1)!}.$$

Multiplying by $n!$ we then get

$$\frac{p}{q} n! - n! \sum_{j=0}^n \frac{1}{j!} < \frac{4}{n+1}$$

that is, $n! \sum_{j=0}^n 1/j!$ and $n! \frac{p}{q}$ being integers for $n \geq q$,

$$\frac{p}{q} = \sum_{j=0}^n \frac{1}{j!} = 0 \quad \forall n \geq \max(3, q),$$

a contradiction. □

6.20 ¶. One can estimate the error better. Show that

$$e - \sum_{j=0}^n \frac{1}{j!} \leq \frac{1}{n!}$$

which yields, for $n = 6$, $2.718055 \dots < e < 2.718287 \dots$ [Hint: Write

$$e - \sum_{j=0}^n \frac{1}{j!} = \sum_{j=n+1}^{\infty} \frac{1}{j!} = \sum_{j=0}^{\infty} \frac{1}{(n+1+j)!}$$

and observe

$$(n+1+j)! = (n+1)!(n+2)(n+3) \cdots (n+j+1) \geq (n+1)!(n+2)^j.]$$

6.3 Series of Nonnegative Terms

The problem of studying the convergence of a series (of real terms) simplifies a great deal if we restrict our attention to series of nonnegative terms. In fact, in this case *the sequence of partial sums is increasing*, therefore it *has a limit that can be finite or infinite*. In particular we can pass to the limit in equalities and inequalities involving the partial sums and we can estimate the sum of the series.

For example, if $a_j \leq b_j$ for all $j \geq p$, then

$$\sum_{j=p}^n a_j \leq \sum_{j=p}^n b_j \quad \forall n \geq p; \quad (6.19)$$

however we cannot take the limit as $n \rightarrow \infty$ until we know the *existence* of the limits $\sum_{j=1}^{\infty} a_j$ e $\sum_{j=1}^{\infty} b_j$. For series of positive terms (or *definitively*¹ nonnegative, or even *definitively* of constant sign) the respective sequences of partial sums are (definitively) monotone, hence the existence of the sum is granted. We therefore can state

6.21 Proposition (Comparison test). Let $\sum_{j=0}^{\infty} a_j$ and $\sum_{j=0}^{\infty} b_j$ be two series of positive terms. Suppose that $a_j \leq b_j$ for all $j \geq p$. Then

- (i) if $\sum_{j=0}^{\infty} b_j$ converges, then $\sum_{j=0}^{\infty} a_j$ converges,
- (ii) if $\sum_{j=0}^{\infty} a_j$ diverges, then $\sum_{j=0}^{\infty} b_j$ diverges.

In both cases

$$\sum_{j=p}^{\infty} a_j \leq \sum_{j=p}^{\infty} b_j. \quad (6.20)$$

¹ We shall say that a predicate $p(n)$ holds *definitively* if there exists \bar{n} such that $p(n)$ holds true for all $n \geq \bar{n}$. Notice that “definitively” is much more than “for infinitely many indices.”

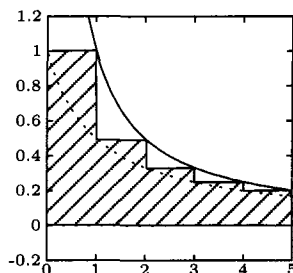


Figure 6.14. The dashed area H_n and the graph of $1/x$, $x > 0$.

6.22 Example. Since

$$\frac{1}{j^2} \leq \frac{2}{j(j+1)}, \quad \forall j \geq 1,$$

we infer

$$\sum_{j=1}^n \frac{1}{j^2} \leq 2 \sum_{j=1}^n \frac{1}{j(j+1)},$$

hence

$$1 \leq \sum_{j=1}^{\infty} \frac{1}{j^2} < 2 \sum_{j=1}^{\infty} \frac{1}{j(j+1)} = 2.$$

Actually, compare 7.79, we have $\sum_{j=1}^{\infty} 1/j^2 = \pi^2/6$.

6.23 Example. Let us show that the series $\sum_{n=1}^{\infty} \log \cos(1/n)$ is convergent.

Observe that all terms of the series are negative, and that $\cos 1 \leq \cos(1/n) < 1 \forall n$. The estimates (see, e.g., Section 5.4 of [GM1])

$$\log x \geq K(x-1), \quad \cos 1 \leq x \leq 1, \quad K := -\frac{\log(\cos 1)}{\cos 1},$$

$$\cos x \geq 1 - \frac{x^2}{2}, \quad x \in \mathbb{R}$$

yield then

$$-\log \cos \frac{1}{n} \leq \frac{K}{2} \frac{1}{n^2},$$

hence

$$-\sum_{n=1}^{\infty} \log \cos \frac{1}{n} \leq \frac{K}{2} \sum_{n=1}^{\infty} \frac{1}{n^2} < +\infty$$

by the comparison test and Example 6.22.

A variant of Proposition 6.21 is

6.24 Proposition (Asymptotic comparison test). Let $\sum_{j=0}^{\infty} a_j$ and $\sum_{j=0}^{\infty} b_j$ be two series of positive terms. Suppose that

$$\frac{a_n}{b_n} \rightarrow L \in \mathbb{R}.$$

(i) If $\sum_{j=0}^{\infty} a_j$ diverges, then $\sum_{j=0}^{\infty} b_j$ diverges.

(ii) If $\sum_{j=0}^{\infty} b_j$ converges, then $\sum_{j=0}^{\infty} a_j$ converges.

Proof. In fact the sequence a_n/b_n is bounded, i.e., $\exists M > 0$ such that $a_n \leq M b_n$ for all n . The comparison test in Proposition 6.21 then yields $\sum_{j=0}^{\infty} a_j \leq M \sum_{j=0}^{\infty} b_j$. \square

a. Series of positive decreasing terms

6.25 Example (Harmonic series, I). An especially relevant series is

$$\sum_{j=1}^{\infty} \frac{1}{j},$$

called the *harmonic series*, since the numbers $1, 1/2, 1/3, \dots, 1/n$ represent the ratio of the lengths of “harmonic” vibrating strings.

There is no closed formula for the partial sums $H_n := \sum_{j=1}^n \frac{1}{j}$. However, it is easily seen that the harmonic series diverges, $H_n \rightarrow \infty$, moreover its partial sums H_n can be estimated by considering the improper integral associated to it.

Let $\varphi: [0, +\infty[\rightarrow \mathbb{R}$ be the piecewise constant function defined by

$$\varphi(x) = \frac{1}{j} \quad \text{if } j-1 \leq x < j.$$

As we have seen, and it is evident (see Figure 6.14),

$$\sum_{j=2}^n 1/j = \int_1^n \varphi(x) dx.$$

On the other hand,

$$\frac{1}{x+1} \leq \varphi(x) \leq \frac{1}{x} \quad \forall x > 0,$$

hence

$$\int_0^n \frac{1}{1+x} dx \leq H_n = 1 + \sum_{j=2}^n \frac{1}{j} \leq 1 + \int_1^n \frac{1}{x} dx,$$

i.e.,

$$\log(1+n) \leq H_n \leq 1 + \log n. \quad (6.21)$$

H_n is therefore asymptotic to $\log n$, $H_n/\log n \rightarrow 1$. In particular H_n tends to infinity quite slowly as $n \rightarrow \infty$, see Figure 6.15.

6.26 Example (Euler–Mascheroni constant). Let us now consider the difference $\gamma_n := H_n - \log n$. From (6.21) we see that $0 < \gamma_n < 1$. Since

$$\gamma_n = 1 + \int_1^n \left(\varphi(x) - \frac{1}{x} \right) dx \quad \text{and} \quad \varphi(x) \leq 1/x \quad \forall x,$$

the sequence $\{\gamma_n\}$ is decreasing and has limit γ . Since $1/x$ is convex, we also deduce for $x \in [j-1, j]$

$$\frac{1}{x} \leq \left(\frac{1}{j} - \frac{1}{j-1} \right) \left(x - \frac{1}{j} \right) + \frac{1}{j},$$

hence

$$\frac{1}{x} - \varphi(x) \leq \left(\frac{1}{j} - \frac{1}{j-1} \right) \left(x - \frac{1}{j} \right);$$

and, integrating on $[-1, n]$,

n	H_n	$H_n / \log n - 1$	$H_n - \log n$
10	2.928968253968254	$3E - 01$	0.626383160974208
30	3.994987130920391	$2E - 01$	0.593789749258235
100	5.187377517639621	$1E - 01$	0.582207331651529
300	6.282663880299502	$1E - 01$	0.578881405643301
1000	7.485470860550343	$8E - 02$	0.577715581568206
3000	8.583749889959170	$7E - 02$	0.577382322308925
10000	9.787606036044345	$6E - 02$	0.577265664068161
30000	10.886184992119919	$6E - 02$	0.577232331475626
100000	12.090146129863282	$5E - 02$	0.577220664893053

Figure 6.15. $H_n = \sum_{j=1}^n 1/j$, $H_n / \log n - 1$ and $H_n - \log n$.

$$\gamma_n = 1 + \int_1^n \left(\varphi(x) - \frac{1}{x} \right) dx \geq 1 - \frac{1}{2} \sum_{j=2}^n \left(\frac{1}{j-1} - \frac{1}{j} \right) = 1 - \frac{1}{2} \left(1 - \frac{1}{n} \right) = \frac{1}{2} + \frac{1}{2n}.$$

In conclusion we can state

Proposition. The partial sums H_n of the harmonic series are asymptotic to $\log n$. Moreover $\{H_n - \log n\}$ is a decreasing sequence with limit $\gamma \in]1/2, 1[$.

The previous constant γ is called the *Euler–Mascheroni constant*. It has an approximate value of 0.57721566..., but it is not known if it is irrational or rational.

From

$$H_n = \log n + \gamma + o(1)$$

we see

$$\frac{H_n}{\log n} = 1 + \frac{\gamma}{\log n} + o\left(\frac{1}{\log n}\right).$$

This explains the slowness of the convergence $H_n / \log n \rightarrow 1$ that one sees in Figure 6.15.

6.27 Example (The harmonic series, II). One can prove that the harmonic series diverges also as follows. Observe that we have

$$\begin{aligned} 1 &= 1 \leq 1, \\ \frac{1}{2} &= \frac{2}{4} < \frac{1}{2} + \frac{1}{3}, \\ \frac{1}{2} &= \frac{4}{8} < \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7}, \\ \frac{1}{2} &= \frac{8}{16} < \frac{1}{8} + \frac{1}{9} + \cdots + \frac{1}{14} + \frac{1}{15}, \\ &\dots \\ \frac{1}{2} &= \frac{2^n}{2^{n+1}} < \sum_{k=2^{n-1}}^{2^n-1} \frac{1}{k}, \end{aligned}$$

thus

$$\frac{n}{2} \leq H_{2^n-1} \quad \text{for all } n,$$

in particular $H_{2^n-1} \rightarrow \infty$. Since $\{H_{2^n-1}\}$ is a subsequence of $\{H_n\}$ and H_n is increasing, we conclude $H_n \rightarrow \infty$.

6.28 Example (The generalized harmonic series, I). Consider the series $\sum_{j=1}^{\infty} 1/n^{\alpha}$, $\alpha \neq 1$, and the piecewise constant function $\varphi : [0, +\infty[\rightarrow \mathbb{R}$ defined by

$$\varphi(x) = \frac{1}{n^{\alpha}} \quad \text{if } n-1 \leq x < n.$$

For $n \geq k \geq 1$ we have

$$\sum_{j=k}^n \frac{1}{j^{\alpha}} = \int_{k-1}^n \varphi(x) dx.$$

Since $1/(x+1)^{\alpha} \leq \varphi(x) \leq 1/x^{\alpha}$ for all $x > 0$, and therefore

$$\int_0^n \frac{1}{(x+1)^{\alpha}} dx \leq \sum_{j=1}^n \frac{1}{j^{\alpha}} \leq 1 + \int_1^n \frac{1}{x^{\alpha}} dx,$$

we conclude

$$\frac{(n+1)^{-\alpha+1} - 1}{-\alpha+1} \leq \sum_{j=1}^n \frac{1}{j^{\alpha}} \leq 1 + \frac{n^{-\alpha+1} - 1}{-\alpha+1}.$$

Therefore we can state

Proposition. The generalized harmonic series, $\sum_{n=1}^{\infty} \frac{1}{n^{\alpha}}$, converges if and only if $\alpha > 1$ and

$$\frac{1}{\alpha-1} \leq \sum_{n=1}^{\infty} \frac{1}{n^{\alpha}} \leq 1 + \frac{1}{\alpha-1}. \quad (6.22)$$

The reasoning in Example 6.27 extends to obtain

6.29 Theorem (Cauchy condensation test). Let $\sum_{j=1}^{\infty} a_j$ be a series of nonnegative and decreasing terms. Then $\sum_{j=1}^{\infty} a_j$ converges if and only if $\sum_{j=0}^{\infty} 2^j a_{2^j}$ converges. Moreover, the following estimates hold:

$$\frac{1}{2} \sum_{j=1}^{\infty} 2^j a_{2^j} \leq \sum_{j=1}^{\infty} a_j \leq \sum_{j=0}^{\infty} 2^j a_{2^j}. \quad (6.23)$$

Proof. By the assumptions made we have

$$\begin{aligned} \frac{1}{2} 2a_2 &\leq a_1 &&\leq a_1, \\ \frac{1}{2} 4a_4 &\leq a_2 + a_3 &&\leq 2a_2, \\ \frac{1}{2} 8a_8 &\leq a_4 + a_5 + a_6 + a_7 &&\leq 4a_4, \\ \frac{1}{2} 16a_{16} &\leq a_8 + a_9 + \cdots + a_{15} &&\leq 8a_8, \\ &\dots && \\ \frac{1}{2} 2^{n+1} a_{2^{n+1}} &\leq \sum_{j=2^n}^{2^{n+1}-1} a_j &< 2^n a_{2^n}. \end{aligned}$$

Summing, we infer

$$\frac{1}{2} \sum_{j=1}^{n+1} 2^j a_{2^j} \leq \sum_{j=1}^{2^{n+1}-1} a_j \leq \sum_{j=0}^n 2^j a_{2^j} \quad (6.24)$$

for all $n \geq 0$. We then conclude

$$\sum_{j=0}^n 2^j a_{2^j} \rightarrow \sum_{j=0}^{\infty} 2^j a_{2^j}, \quad \sum_{j=1}^{n+1} 2^j a_{2^j} \rightarrow \sum_{j=1}^{\infty} 2^j a_{2^j}, \quad \sum_{j=1}^n a_j \rightarrow \sum_{j=1}^{\infty} a_j.$$

On the other hand $\sum_{j=1}^{2^{n+1}-1} a_j$ is a subsequence of $\sum_{j=1}^n a_j$, hence

$$\sum_{j=1}^{2^{n+1}-1} a_j \rightarrow \sum_{j=1}^{\infty} a_j.$$

Passing to the limit in (6.24) we get (6.23), hence the result. \square

6.30 Example (The generalized harmonic series, II). The Cauchy condensation test yields

Proposition. *The generalized harmonic series $\sum_{j=1}^{\infty} \frac{1}{n^\alpha}$ converges if and only if $\alpha > 1$.*

Proof. In fact the assumptions of the Cauchy condensation test theorem are satisfied, therefore $\sum_{n=1}^{\infty} \frac{1}{n^\alpha}$ converges if and only if the geometric series

$$\sum_{j=0}^{\infty} 2^j \frac{1}{2^{\alpha j}} = \sum_{j=0}^{\infty} \left(\frac{1}{2^{\alpha-1}}\right)^j$$

converges. The last converges if and only if $1/2^{\alpha-1} < 1$, that is, $\alpha > 1$. \square

b. The root and ratio tests

Some comparisons are more frequent than others. They lead to rules, called convergence tests. Here we present two of them: *Cauchy's root test* and *d'Alembert's ratio test*.

6.31 Theorem (Root test). *Let $\sum_{n=1}^{\infty} a_n$ be a series of nonnegative terms. Suppose there is a positive constant $K < 1$ and a natural $p \in \mathbb{N}$ such that for all $n \geq p$,*

$$\sqrt[p]{a_n} \leq K < 1.$$

Then $\sum_{n=1}^{\infty} a_n$ converges and

$$\sum_{n=p}^{\infty} a_n \leq \frac{K^p}{1-K}.$$

If there is a positive constant $K > 1$ and a natural $p \in \mathbb{N}$ such that for all $n \geq p$ $\sqrt[p]{a_n} \geq K > 1$, then $\sum_{n=1}^{\infty} a_n$ diverges.

Proof. In fact, for $j \geq p$ we have $0 \leq a_j \leq K^j$ hence

$$\sum_{j=p}^n a_j \leq \sum_{j=p}^n K^j = \sum_{j=0}^{n-p} K^{p+j} \leq \frac{K^p}{1-K}.$$

Passing to the limit as $n \rightarrow \infty$, the claim follows.

Similarly one proves divergence if $\sqrt[n]{a_n} \geq K > 1$. \square

6.32 Proposition (Ratio test). Let $\sum_{n=1}^{\infty} a_n$ be a series of nonnegative terms. Suppose there is a positive constant $K < 1$ and a natural $p \in \mathbb{N}$ such that for all $n \geq p$,

$$\frac{a_{n+1}}{a_n} \leq K < 1.$$

Then $\sum_{n=1}^{\infty} a_n$ converges and

$$\sum_{n=p}^{\infty} a_n \leq \frac{a_p}{1-K}.$$

Suppose there is a positive constant $K > 1$ and a natural $p \in \mathbb{N}$ such that $\frac{a_{n+1}}{a_n} \leq K < 1$ for all $n \geq p$. Then $\sum_{n=1}^{\infty} a_n$ diverges.

Proof. Inductively we find

$$\begin{aligned} a_{p+1} &\leq K a_p, \\ a_{p+2} &\leq K a_{p+1} \leq K^2 a_p, \\ &\dots \\ a_{p+j} &\leq K a_{p+j-1} \leq K^2 a_{p+j-2} \leq \dots \leq K^j a_p, \\ &\dots \end{aligned}$$

hence, summing from p to n ,

$$\sum_{j=p}^n a_j \leq a_p \sum_{j=0}^{n-p} K^j \leq a_p \frac{1}{1-K}.$$

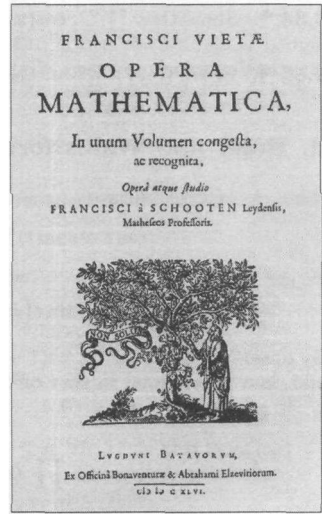
When $n \rightarrow \infty$, we get the result.

Similarly one proves divergence if $a_{n+1}/a_n \geq K > 1$. \square

6.33 Remark. Notice that root and ratio tests are inconclusive if $\sqrt[n]{a_n} \rightarrow 1$ or $a_{n+1}/a_n \rightarrow 1$, as is shown by the generalized harmonic series. Also, because of Example 2.57, whenever the ratio test yields convergence, the root test does, too; whenever the root test is inconclusive, the ratio test is inconclusive, too.



Figure 6.16. François Viète (1540–1603) and the frontispiece of his *Opera Mathematica*.



c. Viète's formula for π

From $\sin x = 2 \sin(x/2) \cos(x/2)$, we infer by induction

$$\sin x = 2^n \sin\left(\frac{x}{2^n}\right) \prod_{k=1}^n \cos\left(\frac{x}{2^k}\right)$$

and, since, $2^n \sin(x/2^n) \rightarrow x$, we find

$$\sin x = x \prod_{k=1}^{\infty} \cos\left(\frac{x}{2^k}\right). \quad (6.25)$$

On the other hand $\cos^2(x/2) = (1 + \cos x)/2$, thus $\cos(x/2) = \sqrt{\frac{1}{2} + \frac{1}{2} \cos x}$ for $x \in [0, \pi/2]$, hence

$$\cos \frac{\pi}{4} = \sqrt{\frac{1}{2}}, \quad \cos \frac{\pi}{8} = \sqrt{\frac{1}{2} + \frac{1}{2} \sqrt{\frac{1}{2}}}, \quad \cos \frac{\pi}{16} = \sqrt{\frac{1}{2} + \frac{1}{2} \sqrt{\frac{1}{2} + \frac{1}{2} \sqrt{\frac{1}{2}}}}, \dots$$

Therefore (6.25) with $x = \pi/2$, yields the *Viète formula*

$$\frac{2}{\pi} = \prod_{n=2}^{\infty} \cos\left(\frac{\pi}{2^n}\right) = \sqrt{\frac{1}{2}} \sqrt{\frac{1}{2} + \frac{1}{2} \sqrt{\frac{1}{2}}} \sqrt{\frac{1}{2} + \frac{1}{2} \sqrt{\frac{1}{2} + \frac{1}{2} \sqrt{\frac{1}{2}}}} \dots$$

Notice that the Viète sequence $\{x_n\}$,

$$x_n := \prod_{k=1}^n \cos\left(\frac{\pi}{2^{k+1}}\right) = \frac{1}{2^n \sin\left(\frac{\pi}{2^{n+1}}\right)}, \quad (6.26)$$

converges exponentially fast to $2/\pi$; in fact we have

$$0 < x_n - \frac{2}{\pi} < C \frac{1}{4^n}. \quad (6.27)$$

6.34 ¶. Show that $\prod_{k=1}^{\infty} \cos(x/2^k)$ converges.

6.35 ¶. Prove (6.27). [*Hint:* Use that $\sin x > x - x^3/3!$ holds for $x \geq 0$.]

d. Euler and Wallis formulas

From de Moivre's formula

$$(\cos t + i \sin t)^k = \cos kt + i \sin kt, \quad t \in \mathbb{R}, k \in \mathbb{Z},$$

we see that

$$\sin kt = \sin t \left(k \cos^{k-1} t - \binom{k}{3} \cos^{k-3} t \sin^2 t + \dots \right).$$

By observing that $\cos^{2n} t = (1 - \sin^2 t)^n$, we readily conclude that $\sin kt$, for $k = 2n + 1$ odd, is a polynomial in $\sin t$ of degree k . Since $\sin((2n + 1)t)$ has $2n + 1$ distinct zeros $t_j := \frac{\pi}{2n+1}j$, $j = -n, \dots, 0, 1, \dots, n$,

$$\sin(2n + 1)t = C \prod_{j=-n}^n (\sin t - \sin t_j) \approx C \sin t \prod_{\substack{j=-n, \dots, n \\ j \neq 0}} (\sin t - \sin t_j).$$

Dividing by C and passing to the limit for $t \rightarrow 0$,

$$C \cdot \prod_{\substack{j=-n, \dots, n \\ j \neq 0}} \sin t_j = 2n + 1$$

and we get

$$\sin(2n + 1)t = (2n + 1) \sin t \prod_{\substack{j=-n, \dots, n \\ j \neq 0}} \left(1 - \frac{\sin t}{\sin t_j} \right) = (2n + 1) \sin t \prod_{j=1}^n \left(1 - \frac{\sin^2 t}{\sin^2 t_j} \right).$$

Finally, replacing $(2n + 1)t$ by x we deduce

$$\sin x = (2n + 1) \sin \left(\frac{x}{2n + 1} \right) \prod_{j=1}^n \left(1 - \frac{\sin^2(x/(2n + 1))}{\sin^2(j\pi/(2n + 1))} \right).$$

When $n \rightarrow \infty$, a “naive” passage to the limit yields

$$\sin x = x \prod_{j=1}^{\infty} \left(1 - \frac{x^2}{j^2 \pi^2} \right), \quad \text{for all } x \neq k\pi, k \in \mathbb{Z} \quad (6.28)$$

that, for $x = \pi/2$, yields, in turn, *Wallis's formula for π* (see Example 2.66),

$$\frac{2}{\pi} = \prod_{n=1}^{\infty} \frac{2n-1}{2n} \frac{2n+1}{2n} \quad \text{or} \quad \frac{\pi}{2} = \prod_{n=1}^{\infty} \frac{2n}{2n-1} \frac{2n}{2n+1}. \quad (6.29)$$

Actually, Euler's formula for \sin is equivalent, for $|x| < \pi$, to

$$\log \frac{\sin x}{x} = \sum_{j=1}^{\infty} \log \left(1 - \frac{x^2}{j^2 \pi^2} \right).$$

Since differentiation term by term is allowed in the series on the right, it turns out to be equivalent also to

$$\cot x - \frac{1}{x} = 2x \sum_{j=1}^{\infty} \frac{1}{j^2 \pi^2 - x^2}, \quad (6.30)$$

known as *Euler's formula for cotangent*.

The natural context of Euler's formulas for \sin and \cot is the *theory of complex functions*. There they arise in a transparent and simple way. As Jacques Hadamard (1865–1963) put it: *Le plus court chemin entre deux énoncés réels passe par le complexe*². Here we prove Euler's formula for $|x| < 1$.

First we state the following theorem that is interesting by itself.

6.36 Theorem (of dominated convergence). Suppose that the double sequence of numbers $a_{j,n}$ is such that

- (i) $a_{j,n} \rightarrow a_j$ as $n \rightarrow \infty$ for all j ,
- (ii) for all n we have $|a_{j,n}| \leq c_j$ with $\sum_{j=1}^{\infty} c_j < \infty$,

Then $\sum_{j=1}^{\infty} a_j$ converges and

$$\sum_{j=1}^{\infty} a_{j,n} \rightarrow \sum_{j=1}^{\infty} a_j \quad \text{when } n \rightarrow \infty.$$

Proof. First observe that, since $a_{j,n} \rightarrow a_j$, and $|a_{n,j}| \leq c_j \forall n, j$, we also have $|a_j| \leq c_j \forall j$, hence $\sum_{j=0}^{\infty} a_j$ converges absolutely.

Fix $\epsilon > 0$ and choose $p = p(\epsilon)$ such that $2 \sum_{j=p+1}^{\infty} c_j < \epsilon$. Then

$$\begin{aligned} \left| \sum_{j=0}^{\infty} a_{j,n} - \sum_{j=0}^{\infty} a_j \right| &\leq \sum_{j=0}^{\infty} |a_{j,n} - a_j| = \sum_{j=0}^p |a_{j,n} - a_j| + \sum_{j=p+1}^{\infty} |a_{j,n} - a_j| \\ &\leq \sum_{j=0}^p |a_{j,n} - a_j| + 2 \sum_{j=p+1}^{\infty} c_j \leq \sum_{j=0}^p |a_{j,n} - a_j| + \epsilon, \end{aligned}$$

hence

$$\limsup_{n \rightarrow \infty} \left| \sum_{j=0}^{\infty} a_{j,n} - \sum_{j=0}^{\infty} a_j \right| \leq \epsilon,$$

and finally the claim, ϵ being arbitrary. \square

Proof of (6.28). Set for $|x| < 2$,

$$a_{j,n} := \begin{cases} \log \left(1 - \frac{\sin^2 \frac{x}{2n+1}}{\sin^2 \frac{j\pi}{2n+1}} \right) & \text{if } j \leq n, \\ 0 & \text{if } j > n, \end{cases}$$

and

$$a_j := \log \left(1 - \frac{x^2}{j^2 \pi^2} \right).$$

Clearly $a_{j,n} \rightarrow a_j$. As $\sin t \geq \frac{2}{\pi} t \forall t \in [0, \pi/2]$, we have

$$\frac{\sin^2 \frac{x}{2n+1}}{\sin^2 \frac{j\pi}{2n+1}} \leq \frac{x^2}{4j^2} \quad \forall j \leq n;$$

consequently

$$|a_{n,j}| \leq -\log \left(1 - \frac{x^2}{4j^2} \right) =: c_j \quad \text{and} \quad \sum_{j=1}^{\infty} c_j < \infty.$$

Applying the theorem of dominated convergence, we conclude

² the shortest path between two real statements is via complex

n	1	2	3	4	5	6	7	8
$a_n = (-1)^n/n$	-1	1/2	-1/3	1/4	-1/5	1/6	-1/7	1/8
a_n^+	0	1/2	0	1/4	0	1/6	0	1/8
a_n^-	1	0	1/3	0	1/5	0	1/7	0
$ a_n $	1	1/2	1/3	1/4	1/5	1/6	1/7	1/8

Figure 6.17. $a_n, a_n^+, a_n^- \in |a_n|$ per $a_n = (-1)^n/n$.

$$\sum_{j=1}^{\infty} \log \left(1 - \frac{\sin^2 \frac{x}{2n+1}}{\sin^2 \frac{j\pi}{2n+1}} \right) \rightarrow \sum_{j=1}^{\infty} \log \left(1 - \frac{x^2}{j^2 \pi^2} \right) \quad \text{as } n \rightarrow \infty,$$

i.e.,

$$\prod_{j=1}^n \left(1 - \frac{\sin^2 \frac{x}{2n+1}}{\sin^2 \frac{j\pi}{2n+1}} \right) \rightarrow \prod_{j=1}^{\infty} \left(1 - \frac{x^2}{j^2 \pi^2} \right)$$

hence Euler's formula for $|x| < 2$. □

6.4 Series of Terms of Arbitrary Sign

In the case where the terms of the series $\sum_{j=0}^{\infty} a_j$ are of arbitrary sign, it is convenient to set

$$a_j^+ := \begin{cases} a_j & \text{if } a_j > 0, \\ 0 & \text{otherwise,} \end{cases} \quad a_j^- := \begin{cases} -a_j & \text{if } a_j < 0, \\ 0 & \text{otherwise.} \end{cases}$$

Trivially $a_j = a_j^+ - a_j^-$ for all $j \geq 0$, hence $\sum_{j=0}^n a_j = \sum_{j=0}^n a_j^+ - \sum_{j=0}^n a_j^-$.

6.37 Proposition. *We have*

- $\sum_{j=0}^{\infty} a_j$ converges if both $\sum_{j=0}^{\infty} a_j^+$ and $\sum_{j=0}^{\infty} a_j^-$ converge,
- $\sum_{j=0}^{\infty} a_j$ diverges to $+\infty$ if $\sum_{j=0}^{\infty} a_j^+$ diverges to $+\infty$ and $\sum_{j=0}^{\infty} a_j^-$ converges,
- $\sum_{j=0}^{\infty} a_j$ diverges to $-\infty$ if $\sum_{j=0}^{\infty} a_j^+$ converges and $\sum_{j=0}^{\infty} a_j^-$ diverges to $+\infty$.

This way the study of the convergence of a generic series of real terms is subsumed to that of a series with nonnegative terms, except in the case where both $\sum_{j=0}^{\infty} a_j^+$ and $\sum_{j=0}^{\infty} a_j^-$ are divergent.

a. Absolute convergence

6.38 Definition. *We say that $\sum_{j=0}^{\infty} a_j$ converges absolutely if the series $\sum_{j=0}^{\infty} |a_j|$ converges.*

6.39 Proposition. If $\sum_{j=0}^{\infty} a_j$ converges absolutely, then it converges and

$$\left| \sum_{j=0}^{\infty} a_j \right| \leq \sum_{j=0}^{\infty} |a_j|. \quad (6.31)$$

Proof. We prove that $\sum_{n=0}^{\infty} |a_n|$ converges if and only if both $\sum_{j=0}^{\infty} a_j^+$ and $\sum_{j=0}^{\infty} a_j^-$ converge. In fact, for $j \geq 0$, a_j^+ and a_j^- are nonnegative and $a_j^+ + a_j^- = |a_j|$, hence $a_j^+, a_j^- \leq |a_j|$. The comparison test yields that the series $\sum_{j=0}^{\infty} a_j^+$ and $\sum_{j=0}^{\infty} a_j^-$ converge, consequently $\sum_{j=0}^{\infty} a_j$ converges. By the triangle inequality, we get

$$\left| \sum_{j=0}^n a_j \right| \leq \sum_{j=0}^n |a_j| \leq \sum_{j=0}^{\infty} |a_j|, \quad \forall n \geq 0,$$

therefore we deduce (6.31) passing to the limit as $n \rightarrow \infty$. \square

b. Series of complex terms

The notion of sum of a series easily extends to series of complex terms. We say that $\sum_{n=0}^{\infty} z_n$ converges (respectively diverges) if the partial sums have finite (respectively infinite) limit. The sum is then defined by

$$\sum_{n=0}^{\infty} z_n := \lim_{n \rightarrow \infty} \sum_{j=0}^n z_j.$$

6.40 Example (Geometric series). For $z \in \mathbb{C}$, $z \neq 1$, we still have

$$\sum_{j=0}^n z^j = (z^{n+1} - 1)(z - 1),$$

therefore $\sum_{n=0}^{\infty} z^n$ converges for $|z| < 1$ and

$$\sum_{n=0}^{\infty} z^n = \frac{1}{1-z} \quad \text{for } |z| < 1.$$

If $|z| > 1$, since

$$|z^{n+1} - 1| > |z|^{n+1} - 1 \rightarrow +\infty \quad \text{as } n \rightarrow \infty,$$

Clearly $\sum_{n=0}^{\infty} z^n$ does not converge if $z = 1$, since $\sum_{j=0}^n z^j = n + 1$. Finally, it can be proved, see Theorem 8.61, that $\sum_{n=0}^{\infty} z^n$ does not converge for any z in the unitary circle $\{z \mid |z| = 1\} \subset \mathbb{C}$, thus concluding that $\sum_{n=0}^{\infty} z^n$ converges if and only if $|z| < 1$ to $\frac{1}{1-z}$.

6.41 ¶. Show that $\sum_{n=0}^{\infty} z^n$ does not converge if $|z| = 1$. [Hint: Assuming $z \neq 1$, let $z := e^{i\theta}$, $\theta \neq 2k\pi$, $k \in \mathbb{Z}$. Show one of the following claims:

- $\{e^{in\theta}\}$ does not have limit as $n \rightarrow \infty$, see Exercise 2.97.
- If $\theta/(2\pi) = \frac{p}{q}$, p, q coprime integers, then $\{e^{in\theta}\}$ has q values.

If $t/(2\pi)$ is irrational, then $\{e^{in\theta}\}$ is dense on the unit circle (see Theorem 8.61). Thus $\{e^{in\theta}\}$ has a limit iff $\theta/(2\pi)$ is integer.

Similarly, we have, see Example 6.5, that $\sum_{n=0}^{\infty} nz^n$ converges if and only if $|z| < 1$ with sum $\sum_{n=0}^{\infty} nz^n = \frac{z}{(z-1)^2}$, $|z| < 1$.

Again, we trivially have,

6.42 Proposition. If $\sum_{n=0}^{\infty} z_n$ converges, then $|z_n| \rightarrow 0$.

Cauchy convergence criterion, Theorem 4.23, yields

6.43 Proposition. The series $\sum_{j=0}^n z_n$, $z_n \in \mathbb{C}$, converges if and only if the sequence of its partial sums is a Cauchy sequence, i.e., iff $\forall \epsilon > 0$ there exists \bar{n} such that $\left| \sum_{j=p}^q z_j \right| < \epsilon$ for all $p, q \geq \bar{n}$.

6.44 Definition. We say that $\sum_{n=0}^{\infty} z_n$, $z_n \in \mathbb{C}$, converges absolutely if the series of nonnegative terms, $\sum_{n=0}^{\infty} |z_n|$, converges.

6.45 Proposition. If the series $\sum_{n=0}^{\infty} z_n$ converges absolutely, then it converges; moreover $\left| \sum_{n=p}^{\infty} z_n \right| \leq \sum_{n=p}^{\infty} |z_n|$ for all p .

6.46 ¶. Prove Proposition 6.45. [Hint: Use (4.12) or Proposition 6.43.]

6.5 Series of Products

In this section we illustrate a few results concerning series of products of complex numbers

$$\sum_{j=0}^{\infty} a_j b_j.$$

In fact, the product structure of the terms helps in giving further results of convergence.

a. Alternating series

6.47 Definition. An alternating series is a series of the type $\sum_{j=0}^{\infty} (-1)^j a_j$, with $a_j \geq 0$ for all $j \geq 0$.

6.48 Theorem (Leibniz test). Let $\sum_{j=0}^{\infty} (-1)^j a_j$, $a_j \geq 0$, be an alternating series. If $\{a_n\}$ is decreasing to zero, then $\sum_{j=0}^{\infty} (-1)^j a_j$ converges and the errors between the sum and the p -th partial sums are estimated by

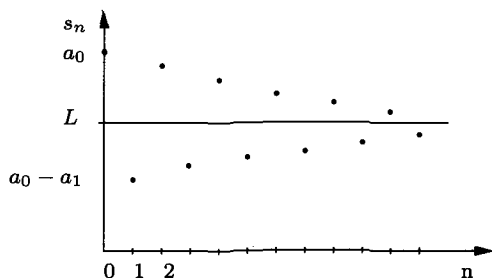


Figure 6.18.

$$a_p - a_{p+1} \leq \left| \sum_{j=p+1}^{\infty} (-1)^j a_j \right| \leq a_p + a_q, \quad \forall p, q, q > p.$$

The inequalities are strict if $\{a_n\}$ is strictly decreasing.

Proof. Let $p < q \in \mathbb{N}$. It is easily seen using the assumptions that

$$\left| \sum_{j=p}^q (-1)^j a_j \right| \leq a_p + a_q. \quad (6.32)$$

Given $\epsilon > 0$, we can find \bar{n} such that $|a_p| < \epsilon$ for all $p \geq \bar{n}$; (6.32) then yields

$$\left| \sum_{j=p}^q (-1)^j a_j \right| < 2\epsilon \quad \text{for all } q > p \geq \bar{n},$$

i.e., the sequence of partial sums of $\sum_{j=0}^{\infty} (-1)^j a_j$ is a Cauchy sequence, therefore convergent. The estimate easily follows from (6.32) letting q tend to infinity. \square

An alternative proof of Theorem 6.48. For $n \in \mathbb{N}$ set $s_n := \sum_{j=0}^n (-1)^j a_j$.

- (i) The subsequence s_{2n} , $n \geq 0$, of $\{s_n\}$, is decreasing and bounded below: in fact,

$$s_{2n+2} = s_{2n} - a_{2n+1} + a_{2n+2} \leq s_{2n}, \quad \forall n \geq 0,$$

$$s_{2n} = a_0 - a_1 + (a_2 - a_3) + \cdots + (a_{2n-2} - a_{2n-1}) + a_{2n} \geq a_0 - a_1,$$

since $\{a_n\}$ is decreasing.

- (ii) The subsequence s_{2n+1} , $n \geq 0$, of $\{s_n\}$, is increasing and bounded above: in fact,

$$s_{2n+1} = s_{2n-1} + a_{2n} - a_{2n+1} \geq s_{2n-1}, \quad \forall n \geq 1,$$

$$s_{2n+1} = a_0 - (a_1 - a_2) - (a_3 - a_4) + \cdots - (a_{2n-1} - a_{2n}) + a_{2n} - a_{2n+1} \leq a_0,$$

since a_n is decreasing.

- (iii) Finally

$$s_{2n+1} = s_{2n} - a_{2n+1} \leq s_{2n}, \quad \forall n,$$

$$s_{2n+1} - s_{2n} = -a_{2n+1} \rightarrow 0 \text{ for } n \rightarrow \infty.$$

By (i) and (ii)

$$s_{2n} \rightarrow L \in \mathbb{R}, \quad s_{2n+1} \rightarrow M \in \mathbb{R}$$

and by (iii)

$$s_{2n} - s_{2n+1} \rightarrow L - M = 0,$$

i.e., s_{2n} and s_{2n+1} have the same limit L . Since the indices of s_{2n} (the even integers) and the indices of s_{2n+1} (the odd integers) exhaust all integers, we then conclude that $s_n \rightarrow L$. Also

$$s_{2n+1} \leq L \leq s_{2n}, \quad \forall n \geq 0, \quad (6.33)$$

hence

$$\begin{aligned} a_{2n+1} - a_{2n+2} &\leq s_{2n} - L \leq s_{2n} - s_{2n+1} = a_{2n+1}, \\ a_{2n+2} - a_{2n+3} &\leq L - s_{2n+1} \leq s_{2n+2} - s_{2n+1} = a_{2n+2}, \end{aligned}$$

that is,

$$\left| \sum_{j=n}^{\infty} (-1)^j a_j \right| = |L - s_{n-1}| \leq a_n \quad \text{for all } n \in \mathbb{N}.$$

□

6.49 Remark. We notice that there exist sequences $a_n \geq 0$, $a_n \rightarrow 0$, for which $\sum_{j=0}^{\infty} (-1)^j a_j$ does not converge, i.e., we cannot omit the assumption that $\{a_n\}$ is decreasing. An example is given by the series of alternating terms

$$(-1)^n a_n := (-1)^n \frac{1}{\sqrt{n}} + \frac{1}{n}$$

whose partial sums are given by

$$\sum_{j=1}^n (-1)^j a_j = \sum_{j=1}^n \frac{(-1)^j}{\sqrt{j}} + \sum_{j=1}^n \frac{1}{j}.$$

The series $\sum_{j=1}^{\infty} (-1)^j / \sqrt{j}$ converges by the Leibniz test, while $\sum_{j=1}^{\infty} \frac{1}{j}$ diverges. Therefore $\sum_{j=1}^{\infty} (-1)^j a_j = +\infty$.

6.50 Remark. The example in Remark 6.49 shows also that the asymptotic comparison test is not valid for series of terms of arbitrary sign. In fact,

$$\frac{\frac{(-1)^n}{\sqrt{n}} + \frac{1}{n}}{\frac{(-1)^n}{\sqrt{n}}} = 1 + \frac{(-1)^n}{\sqrt{n}} \rightarrow 1 \text{ as } n \rightarrow \infty$$

and the series $\sum_{n=1}^{\infty} (-1)^n / \sqrt{n}$ converges, but $\sum_{n=1}^{\infty} (-1)^n a_n$ diverges.

b. Summation by parts

6.51 Proposition (Summation by parts). Let $\{a_n\}$, $\{b_n\}$, be two sequences in \mathbb{C} and let $B_n := \sum_{j=0}^n b_j$, $n \geq 0$, and $B_{-1} = 0$. For arbitrary $p, q \in \mathbb{N}$, $0 \leq p \leq q$, we have

$$\sum_{j=p}^q a_j b_j = \sum_{j=p}^{q-1} (B_j - B_{p-1})(a_j - a_{j+1}) + a_q (B_q - B_{p-1}). \quad (6.34)$$

In particular

$$\left| \sum_{j=p}^q a_j b_j \right| \leq \sup_{p \leq j \leq q} |B_j - B_{p-1}| \left\{ \sum_{j=p}^{q-1} |a_j - a_{j+1}| + |a_q| \right\}. \quad (6.35)$$

Proof. Set $C_j := B_j - B_{p-1}$ so that $C_{p-1} = 0$, and, $b_j = C_j - C_{j-1}$. Then

$$\begin{aligned} \sum_{j=p}^q a_j b_j &= \sum_{j=p}^q a_j (C_j - C_{j-1}) = \sum_{j=p}^q a_j C_j - \sum_{j=p-1}^{q-1} a_{j+1} C_j \\ &= a_q C_q + \sum_{j=p}^{q-1} C_j (a_j - a_{j+1}) - a_p C_{p-1} = a_q C_q + \sum_{j=p}^{q-1} C_j (a_j - a_{j+1}), \end{aligned}$$

that is (6.34). By (6.34) and the triangle inequality, we finally infer

$$\begin{aligned} \left| \sum_{j=p}^q a_j b_j \right| &= \left| \sum_{j=p}^{q-1} \left((B_j - B_{p-1})(a_j - a_{j+1}) \right) + a_q (B_q - B_{p-1}) \right| \\ &\leq \sum_{j=p}^{q-1} \left(|B_j - B_{p-1}| |a_j - a_{j+1}| \right) + |a_q| |B_q - B_{p-1}| \\ &\leq \sup_{p \leq j \leq q} |B_j - B_{p-1}| \left\{ \sum_{j=p}^{q-1} |a_j - a_{j+1}| + |a_q| \right\}. \end{aligned}$$

□

c. Sequences of bounded total variation

6.52 Definition. We say that the sequence $\{a_n\} \subset \mathbb{C}$ has bounded total variation if $\sum_{j=0}^{\infty} |a_j - a_{j+1}|$ converges.

6.53 Example. Let $\sum_{j=0}^{\infty} a_j$ converge absolutely. Then $\{a_n\}$ has bounded total variation since for any $p \geq 0$ we have $\sum_{j=0}^p |a_j - a_{j+1}| \leq 2 \sum_{j=0}^p (|a_j| + |a_{j+1}|)$.

Notice that the sequence $\{(-1)^n/n\}$, which converges to zero, does not have bounded total variation since

$$\sum_{j=1}^{\infty} \left| \frac{(-1)^j}{j} - \frac{(-1)^{j+1}}{j+1} \right| = \sum_{j=1}^{\infty} \left(\frac{1}{j} + \frac{1}{j+1} \right) = +\infty.$$

The following proposition collects a few facts concerning sequences with bounded total variation.

6.54 Proposition. *We have*

- (i) *If $\{a_n\} \subset \mathbb{C}$ has bounded total variation, then $\{a_n\}$ converges, $a_n \rightarrow \ell$, and*

$$|a_p - \ell| \leq \sum_{j=p}^{\infty} |a_j - a_{j+1}|, \quad \forall p \geq 0.$$

- (ii) *Every real, monotone and bounded sequence $\{a_n\}$ has bounded total variation and $\sum_{j=0}^{\infty} |a_j - a_{j+1}| = |a_0 - \lim_{j \rightarrow \infty} a_j|$.*

Proof. (i) For $p < q$ we have

$$|a_q - a_p| = \left| \sum_{j=p}^{q-1} (a_j - a_{j+1}) \right| \leq \sum_{j=p}^{q-1} |a_j - a_{j+1}|.$$

If $\sum_{j=0}^{\infty} |a_j - a_{j+1}|$ converges, then the sequence $\sum_{j=1}^k |a_j - a_{j+1}|$, $k \in \mathbb{N}$, is a Cauchy sequence, i.e., $\forall \epsilon > 0 \exists \bar{p}$ such that

$$\sum_{j=p}^{q-1} |a_j - a_{j+1}| < \epsilon$$

for $p, q \geq \bar{p}$. From

$$|a_q - a_p| = \left| \sum_{j=p}^{q-1} (a_j - a_{j+1}) \right| \leq \sum_{j=p}^{q-1} |a_j - a_{j+1}| \quad (6.36)$$

we then infer $|a_p - a_q| \leq \epsilon$ for $p, q \geq \bar{p}$, i.e., $\{a_n\}$ is a Cauchy sequence, hence $a_n \rightarrow \ell$.

- (ii) For instance, assume $\{a_n\}$ is increasing. Then

$$\sum_{j=0}^n |a_j - a_{j+1}| = \sum_{j=0}^n (a_j - a_{j+1}) = a_0 - a_{n+1},$$

and the conclusion follows when $n \rightarrow \infty$. □



Figure 6.19. Lejeune Dirichlet (1805–1859) and Niels Henrik Abel (1802–1829).

d. Dirichlet and Abel theorems

6.55 Theorem (Dirichlet). Let $\{a_n\}$ and $\{b_n\}$ be two complex sequences. Suppose that

- (i) $\{a_n\}$ has bounded total variation and $a_n \rightarrow 0$,
- (ii) the partial sums of $\{b_n\}$, $B_n := \sum_{j=0}^n b_j$, are bounded, $|B_n| \leq M \in \mathbb{R}$, $\forall n \geq 0$.

Then $\sum_{j=0}^{\infty} a_j b_j$ converges and

$$\left| \sum_{j=p}^{\infty} a_j b_j \right| \leq 2M \sum_{j=p}^{\infty} |a_j - a_{j+1}| \quad \text{for all } p \in \mathbb{N}.$$

Proof. Given $\epsilon > 0$, from the assumptions on $\{a_n\}$ we infer that there exists \bar{p} such that

$$\sum_{j=p}^q |a_j - a_{j+1}| + |a_q| \leq \epsilon$$

for all p, q $q \geq p \geq \bar{p}$. This, together with (6.35) and the assumption on $\{b_n\}$ yields

$$\left| \sum_{j=p}^q a_j b_j \right| \leq \sup_{p \leq j \leq q} |B_j - B_{p-1}| \epsilon \leq 2M \epsilon,$$

that is, $\{\sum_{j=0}^n a_j b_j\}$ is a Cauchy sequence, hence converges. Letting $q \rightarrow \infty$ in (6.35) we finally get the estimate. \square

6.56 Remark. Notice that the Dirichlet test, Theorem 6.55, is an extension of the Leibniz test for alternating series.

6.57 Theorem (Abel). Let $\{a_n\}$ and $\{b_n\}$ be two complex sequences. Suppose that

- (i) $\{a_n\}$ has bounded total variation,
(ii) $\sum_{j=0}^{\infty} b_j$ converges.

Then $\sum_{j=0}^{\infty} a_j b_j$ converges and

$$\left| \sum_{j=p}^{\infty} a_j b_j \right| \leq \sup_{j \geq p} |B_j - B_{p-1}| \left\{ 2 \sum_{j=p}^{\infty} |a_j - a_{j+1}| + |a_p| \right\} \quad \text{for all } p \geq 1.$$

Proof. By (ii) $B_n := \sum_{j=0}^n b_j$ is a Cauchy sequence: $\forall \epsilon > 0 \exists \bar{p} \in \mathbb{N}$ such that $|B_j - B_{p-1}| \leq \epsilon$ for all $j \geq \bar{p}$. By (i) and (ii) of Proposition 6.54, $\{a_n\}$ is convergent, hence bounded, $|a_n| \leq M \in \mathbb{R} \forall n \geq 0$. Therefore we deduce from (6.35)

$$\left| \sum_{j=p}^q a_j b_j \right| \leq \epsilon \left\{ \sum_{j=p}^q |a_j - a_{j+1}| + a_q \right\}. \quad (6.37)$$

In particular the sequence of partial sums of $\sum_{j=0}^{\infty} a_j b_j$ is a Cauchy sequence, hence converges. Finally the estimate follows letting $q \rightarrow \infty$ in (6.37), taking into account

$$|a_q| \leq |a_p| + \sum_{j=p}^{q-1} |a_j - a_{j+1}|.$$

□

6.6 Products of Series

Let $P(x) = \sum_{j=0}^p a_j x^j$ and $Q(x) = \sum_{j=0}^q b_j x^j$ be two polynomials. Recall that

$$P(x)Q(x) = \sum_{k=0}^{p+q} \left(\sum_{i+j=k} a_i b_j \right) x^k \quad (6.38)$$

where we have set $a_i = b_j = 0$ for all i, j such that $p < i \leq p+q$ and $q < j \leq p+q$.

6.58 Definition. Given two sequences $a := \{a_n\}$ and $b := \{b_n\}$, the product of convolution of a and b , denoted $a * b$, is the sequence defined as

$$(a * b)_n := \sum_{i+j=n} a_i b_j = \sum_{j=0}^n a_j b_{n-j}.$$

6.59 Example. The product of convolution is extremely useful in operating with sequences. We give a few examples. If $\delta_{k,n}$ is *Kronecker's symbol*

$$\delta_{k,n} = \begin{cases} 1 & \text{if } k = n, \\ 0 & \text{if } k \neq n, \end{cases}$$

and \mathbf{e}_k is the sequence defined by

$$\mathbf{e}_k := \{\delta_{k,n}\} = \{0, 0, 0, 0, \dots, 1, 0, 0, \dots\},$$

we have

$$(a * \mathbf{e}_k)_n = \begin{cases} 0 & \text{if } n < k, \\ a_{n-k} & \text{if } n \geq k, \end{cases}$$

that is, the values of $a * \mathbf{e}_k$ are the values of $\{a_n\}$ shifted k positions on the right,

$$\begin{aligned} a &= \{a_0, a_1, a_2, \dots, a_n, \dots\} \\ a * \mathbf{e}_k &= \{0, 0, 0, \dots, 0, a_0, a_1, a_2, \dots, a_n, \dots\}. \end{aligned}$$

Similarly, if $b = \{1/2, 1/2, 0, 0, 0, \dots\}$, then

$$(a * b)_n = \begin{cases} a_0/2 & \text{if } n = 0, \\ (a_n + a_{n-1})/2 & \text{if } n \geq 1. \end{cases}$$

If $b = \{1, 1, 1, \dots, 1, \dots\}$, then

$$(a * b)_n = \sum_{k=0}^n a_k \quad \forall n.$$

In terms of product of convolution, (6.38) can be restated as: *the coefficients of $P(x)Q(x)$ are the product of convolution of the coefficients of $P(x)$ and $Q(x)$* , or, better, of the sequences

$$\begin{aligned} a &= \{a_0, a_1, a_2, \dots, a_p, 0, 0, 0, \dots\} \\ b &= \{b_0, b_1, b_2, \dots, b_q, 0, 0, 0, \dots\}, \end{aligned}$$

$$P(x)Q(x) = \sum_{k=0}^{p+q} (a * b)_k x^k.$$

More generally, we have the following.

6.60 Theorem. Let $\sum_{j=0}^{\infty} a_j$ and $\sum_{j=0}^{\infty} b_j$ be absolutely convergent. Then $\sum_{j=0}^{\infty} (a * b)_j$ is absolutely convergent and

$$\sum_{j=0}^{\infty} (a * b)_j = \left(\sum_{j=0}^{\infty} a_j \right) \left(\sum_{j=0}^{\infty} b_j \right).$$

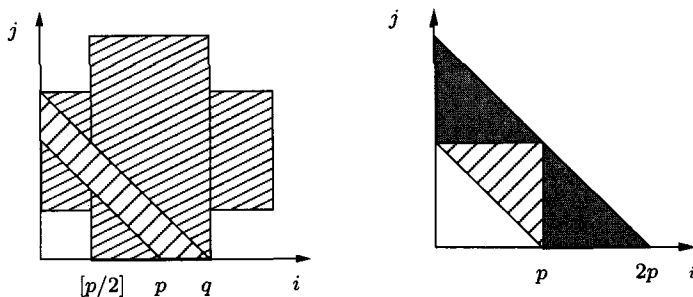


Figure 6.20.

Proof. (i) Set $A := \sum_{j=0}^{\infty} |a_j|$ and $B := \sum_{j=0}^{\infty} |b_j|$. If $q > p$ and $n := [p/2]$ we have

$$\begin{aligned} \left| \sum_{k=p}^q (a * b)_k \right| &\leq \sum_{k=p}^q \left| \sum_{i+j=k} a_i b_j \right| \leq \sum_{p \leq i+j \leq q} |a_i| |b_j| \\ &\leq \sum_{i=n}^q |a_i| \sum_{j=0}^q |b_j| + \sum_{j=n}^q |b_j| \sum_{i=0}^q |a_i| \leq B \sum_{i=n}^q |a_i| + A \sum_{j=n}^q |b_j|. \end{aligned} \quad (6.39)$$

Given $\epsilon > 0$ we find \bar{p} such that $\sum_{i=n}^q |a_i| < \epsilon$ and $\sum_{j=n}^q |b_j| < \epsilon$ for all $q > n \geq \bar{p}$, consequently, on account of (6.39)

$$\left| \sum_{j=p}^q (a * b)_j \right| \leq (A + B)\epsilon \quad \text{for all } q > p > 2\bar{p}.$$

Therefore the sequence of the partial sums of $\sum_{j=0}^{\infty} |(a * b)_j|$ is a Cauchy sequence, hence converges, that is, $\sum_{j=0}^{\infty} (a * b)_j$ converges absolutely.

For $p > 0$ we also have, similarly to (6.39),

$$\begin{aligned} \left| \sum_{k=0}^p \sum_{i+j=k} a_j b_j - \sum_{j=0}^p a_j \sum_{j=0}^p b_j \right| &= \left| \sum_{\substack{i \leq p, j \leq p \\ i+j > p}} a_i b_j \right| \\ &\leq \sum_{p < i+j \leq 2p} |a_i| |b_j| \leq A \sum_{j=n}^{\infty} |b_j| + B \sum_{j=n}^{\infty} |a_j| \end{aligned}$$

where $n := [p/2]$. Passing to the limit as $p \rightarrow \infty$, we get the result. \square

6.61 Remark. There seems to be no known necessary and sufficient condition for the convergence of the series of products. However, one can show

(i) ABEL. If $\sum_{j=0}^{\infty} a_j$, $\sum_{j=0}^{\infty} b_j$ and $\sum_{j=0}^{\infty} (a * b)_j$ converges, then

$$\sum_{j=0}^{\infty} (a * b)_j = \left(\sum_{j=0}^{\infty} a_j \right) \left(\sum_{j=0}^{\infty} b_j \right),$$

see Theorem 7.33.

- (ii) **MERTENS.** If $\sum_{j=0}^{\infty} a_j$ converges and $\sum_{j=0}^{\infty} b_j$ is absolutely convergent, then $\sum_{j=0}^{\infty} (a * b)_j$ converges and by (i), we have $\sum_{j=0}^{\infty} (a * b)_j = \left(\sum_{j=0}^{\infty} a_j \right) \left(\sum_{j=0}^{\infty} b_j \right)$.
- (iii) **HARDY.** If $\sum_{j=0}^{\infty} a_j$ and $\sum_{j=0}^{\infty} b_j$ converge and the sequences $\{na_n\}$ and $\{nb_n\}$ are bounded, then $\sum_{j=0}^{\infty} (a * b)_j$ converges.

6.7 Rearrangements

Given an *ordered enumeration of numbers*, we defined their sum in the previous section. This notion is useful to define the sum of a denumerable set of numbers, but a priori such a sum depends on the order in which they are listed.

A sequence $\{b_n\}$ is a *rearrangement* of $\{a_n\}$ if it contains the same elements of $\{a_n\}$ listed in a different order. More precisely

6.62 Definition. We say that $\{b_n\}$ is a rearrangement of $d\{a_n\}$ if there is a bijective map $k : \mathbb{N} \rightarrow \mathbb{N}$ such that

$$b_n = a_{k_n} \quad \forall n.$$

We say that $\sum_{n=0}^{\infty} b_n$ is a rearrangement of $\sum_{n=0}^{\infty} a_n$.

6.63 Theorem (Dirichlet). Suppose that $\sum_{n=0}^{\infty} a_n^+$ and $\sum_{n=0}^{\infty} a_n^-$ are not both divergent. Then every rearrangement

$$\sum_{n=0}^{\infty} b_n$$

of $\sum_{n=0}^{\infty} a_n$ has the same sum of $\sum_{n=0}^{\infty} a_n$,

$$\sum_{n=0}^{\infty} b_n = \sum_{n=0}^{\infty} a_n = \sum_{n=0}^{\infty} a_n^+ - \sum_{n=0}^{\infty} a_n^-.$$

In particular, in the case of series of nonnegative terms or of absolutely convergent series, the sum is independent of the order of the addends.

Proof. It suffices to prove the theorem in the case of series of nonnegative terms. Let $k : \mathbb{N} \rightarrow \mathbb{N}$ be a map which reorders $\{a_n\}$, set $b_n := a_{k_n}$, and let s_n and σ_n be the n -th partial sums respectively of $\sum_{j=0}^{\infty} a_j$ and $\sum_{j=0}^{\infty} b_j$. Being that $a_n, b_n \geq 0$, we have

$$s_n \rightarrow S := \sum_{j=0}^{\infty} a_j, \quad \sigma_n \rightarrow \Sigma := \sum_{j=0}^{\infty} b_j.$$

Since for every n

$$\sigma_n = b_0 + b_1 + \cdots + b_n = a_{k_0} + a_{k_1} + \cdots + a_{k_n} \leq a_0 + a_1 + \cdots + a_{\max(k_1, \dots, k_n)} \leq S,$$

we deduce $\Sigma \leq S$. Being that $\sum_{j=0}^{\infty} a_j$ is a rearrangement of $\sum_{j=0}^{\infty} b_j$, we also have $S \leq \Sigma$. In conclusion $S = \Sigma$. \square

However, this is not true anymore if

$$\sum_{j=0}^{\infty} a_j^+ = \sum_{j=0}^{\infty} a_j^- = +\infty.$$

6.64 Example. Consider the series

$$\begin{aligned} \sum_{j=0}^{\infty} a_j &:= \left(1 + \frac{1}{3} - \frac{1}{2}\right) + \left(\frac{1}{5} + \frac{1}{7} - \frac{1}{4}\right) + \left(\frac{1}{9} + \frac{1}{11} - \frac{1}{6}\right) + \cdots \\ &\quad + \left(\frac{1}{4j-3} + \frac{1}{4j-1} - \frac{1}{2j}\right) + \cdots \end{aligned}$$

of positive terms, which is convergent since

$$a_j := \frac{1}{4j-3} + \frac{1}{4j-1} - \frac{1}{2j} = \frac{8j-3}{2j(4j-3)4j-1} < \frac{1}{2j^2}.$$

Notice also that

$$\frac{11}{12} < a_1 + a_2 < \sum_{j=0}^{\infty} a_j < +\infty.$$

Removing the parentheses we get

$$\sum_{j=0}^{\infty} b_j = 1 + \frac{1}{3} - \frac{1}{2} + \frac{1}{5} + \frac{1}{7} - \frac{1}{4} + \cdots,$$

which is a rearrangement of the series

$$\sum_{j=0}^{\infty} c_j = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots = \sum_{j=0}^{\infty} \frac{(-1)^n}{n+1},$$

which converges by the Leibniz test to a number L between $1 - 1/2 = 1/2$ and $1 - 1/2 + 1/3 = 5/6 < 11/12$. The sums of the two series $\sum_{j=0}^{\infty} a_j$ and $\sum_{j=0}^{\infty} b_j$ are therefore different.

Of course not all rearrangements change the sums. For instance, a sum does not change if we reorder only a finite number of terms. In general, however, we have

6.65 Theorem (Dini–Riemann). Suppose that $\{a_n\}$ is a sequence which converges to zero and for which

$$\sum_{j=0}^{\infty} a_j^+ = \sum_{j=0}^{\infty} a_j^- = +\infty.$$

Then

- (i) for any $\ell \in \overline{\mathbb{R}}$ there exists a rearrangement $\{b_n\}$ of $\{a_n\}$ such that $\sum_{j=0}^{\infty} b_j = \ell$.
- (ii) There exists a rearrangement $\{b_n\}$ of $\{a_n\}$ such that $\sum_{j=0}^{\infty} b_j$ is indeterminate.

We give an idea of how to define a rearrangement with sum ℓ , $0 \leq \ell \in \mathbb{R}$. We begin by adding in order the nonnegative terms a_j^+ , until we exceed ℓ (this is possible since $\sum_{j=0}^{\infty} a_j^+ = +\infty$). At this point we start to add the negative terms $-a_j^-$ until the sum falls below ℓ (and this is possible since $\sum_{j=0}^{\infty} a_j^- = +\infty$), and then we repeat the procedure. The partial sums of the rearrangement constructed this way oscillate around ℓ , and actually converge to ℓ since $a_j \rightarrow 0$. In other words the idea of the proof is: if one is allowed for unlimited credits and debts and to freely defer takings and payments, then one can decide the threshold of one's own richness or poverty.

We conclude this section by stating a simple consequence of the above concerning double series.

6.66 Proposition. *Given a double sequence $\{a_{ij}\}$ $i, j = 0, 1, 2, \dots$, suppose that $\sum_{j=0}^{\infty} a_{ij}$ is absolutely convergent, and if*

$$b_i := \sum_{j=0}^{\infty} |a_{ij}|, \quad i = 0, 1, 2, \dots,$$

$\sum_{i=0}^{\infty} b_i$ converges. Then

$$\sum_{i=0}^{\infty} \sum_{j=0}^{\infty} a_{ij} = \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} a_{ij}. \quad (6.40)$$

6.67 ¶¶. Prove Proposition 6.66. Show that (6.40) does not hold in general if we only require that $\sum_{j=0}^{\infty} a_{ij}$ converges, that $\sum_{i=0}^{\infty} b_i$ converges, where this time $b_i := \sum_{j=0}^{\infty} a_{ij}$.

6.8 Summing Up

Definitions and basic facts

Given a sequence $\{a_n\}$ of complex numbers, define for every $n \geq 0$ $s_n := \sum_{j=0}^n a_j$. The sequence $\{s_n\}$ is called the *series of partial sums of $\{a_n\}$* and denoted by $\sum_{j=0}^{\infty} a_j$. A series is said to be *convergent* if $\{s_n\}$ has a finite limit, *divergent* if $\{s_n\}$ has an infinite limit, and *indeterminate* if $\{s_n\}$ has no limit.

- If $\sum_{j=0}^{\infty} a_j$ converges, then $a_n \rightarrow 0$ as $n \rightarrow \infty$. The converse is false, in general.
- Given a series $\sum_{j=0}^{\infty} a_j$ of real terms, denote by $\varphi(x)$ the piecewise constant function defined by

$$\varphi(x) = a_j \quad \text{if } j \leq x < j+1.$$

Then trivially

$$\sum_{j=0}^n a_j = \int_0^{n+1} \varphi(x) dx \quad \text{for all } n.$$

Thus partial sums of a series are integrals. This way, the comparison test for integrals becomes a means to estimate the partial sums of a series. Moreover, $\sum_{j=0}^{\infty} a_j$ converges, diverges, or is indeterminate if and only if $\int_0^x \varphi(s) ds$ respectively converges, diverges or is indeterminate as $x \rightarrow \infty$.

Series with nonnegative terms

In this case the sequence $\{s_n\}$ of the partial sums, $s_n := \sum_{j=0}^n a_j$, $a_j \in \mathbb{R}$, $a_j \geq 0 \forall j$, is monotonically increasing, hence $\sum_{j=0}^{\infty} a_j$ either converges or diverges. Consequently, the *comparison test*, Proposition 6.21, and the *asymptotic comparison test*, Proposition 6.24, hold.

The family of the generalized harmonic series

$$\sum_{n=1}^{\infty} \frac{1}{n^{\alpha}}$$

is useful when using the comparison tests. They converge if and only if $\alpha > 1$, and in this case

$$\frac{1}{\alpha - 1} \leq \sum_{n=1}^{\infty} \frac{1}{n^{\alpha}} \leq 1 + \frac{1}{\alpha - 1}.$$

The *harmonic series* $\sum_{n=1}^{\infty} \frac{1}{n}$ diverges. Moreover its partial sums are asymptotic to $\log n$, since we have

$$\log(1+n) < \sum_{j=1}^n \frac{1}{j} < 1 + \log n.$$

There are some other useful tests for convergence,

- CAUCHY'S TEST. Let $\{a_n\}$ be nonnegative and decreasing. Then $\sum_{j=1}^{\infty} a_j$ converges if and only if $\sum_{j=0}^{\infty} 2^j a_{2^j}$ converges. In this case

$$\frac{1}{2} \sum_{j=1}^{\infty} 2^j a_{2^j} \leq \sum_{j=1}^{\infty} a_j \leq \sum_{j=0}^{\infty} 2^j a_{2^j}.$$

- ROOT TEST. Let $\sum_{n=1}^{\infty} a_n$ be a series with nonnegative terms.

- (i) Suppose that there exist $K < 1$ and $p \in \mathbb{N}$ such that $\sqrt[p]{a_n} \leq K < 1$, for all $n \geq p$, then $\sum_{n=1}^{\infty} a_n$ converges and

$$\sum_{n=p}^{\infty} a_n \leq \frac{K^p}{1-K}.$$

- (ii) Suppose that there exist $K > 1$ and $p \in \mathbb{N}$ such that $\sqrt[p]{a_n} \geq K > 1$, for all $n \geq p$, then $\sum_{n=1}^{\infty} a_n$ diverges.

- RATIO TEST. Let $\sum_{n=1}^{\infty} a_n$ be a series with positive terms.

- (i) Suppose that there exist $K < 1$ and $p \in \mathbb{N}$ such that $a_{n+1}/a_n \leq K < 1$, for all $n \geq p$, then $\sum_{n=1}^{\infty} a_n$ converges and

$$\sum_{n=p}^{\infty} a_n \leq \frac{a_p}{1-K}.$$

- (ii) Suppose that there exist $K > 1$ and $p \in \mathbb{N}$ such that $a_{n+1}/a_n \geq K > 1$, for all $n \geq p$, then $\sum_{n=1}^{\infty} a_n$ diverges.

Actually the root and the ratio tests essentially reduce to a comparison test with the *geometric series* $\sum_{j=0}^{\infty} K^j$ for which we have

$$\sum_{j=0}^{\infty} K^j = \begin{cases} +\infty & \text{if } K \geq 1, \\ \frac{1}{1-K} & \text{if } |K| < 1, \\ \text{is indeterminate} & \text{if } K \leq -1. \end{cases}$$

Absolute convergence

We say that $\sum_{j=0}^{\infty} a_j$ *converges absolutely* if $\sum_{j=0}^{\infty} |a_j|$ converges. In case $\{a_n\}$ is a sequence of reals, set

$$a_n^+ := \max(a_n, 0), \quad a_n^- := -\min(a_n, 0).$$

- if $\sum_{j=0}^{\infty} a_j$, $a_j \in \mathbb{C}$, converges absolutely, then $\sum_{j=0}^{\infty} a_j$ converges and $\left| \sum_{j=0}^{\infty} a_j \right| \leq \sum_{j=0}^{\infty} |a_j|$.
- if $\sum_{j=0}^{\infty} a_j$ is a series with real terms, then it converges absolutely if and only if both $\sum_{j=0}^{\infty} a_j^+$ and $\sum_{j=0}^{\infty} a_j^-$ converge, since $0 < a_n^+, a_n^- \leq |a_n| = a_n^+ + a_n^-$ and $a_n = a_n^+ - a_n^-$.
- the complex series $\sum_{j=0}^{\infty} a_j$ converges absolutely if and only if the four series with nonnegative terms $\sum_{j=0}^{\infty} \Re(a_j)^+$, $\sum_{j=0}^{\infty} \Re(a_j)^-$, $\sum_{j=0}^{\infty} \Im(a_j)^+$ and $\sum_{j=0}^{\infty} \Im(a_j)^-$ converge.

Series of products

An alternating series is a series of the type $\sum_{j=0}^{\infty} (-1)^j a_j$, where $a_j \geq 0$ for all j .

- **LEIBNIZ TEST.** Assume that $\{a_n\}$ is monotonically decreasing to zero. Then the alternating series $\sum_{j=0}^{\infty} (-1)^j a_j$ converges. Moreover we have the following estimate for the error between the sum of the series and the n -th partial sum:

$$\left| \sum_{j=n+1}^{\infty} (-1)^j a_j \right| \leq a_{n+1}.$$

The assumption that $\{a_n\}$ is decreasing cannot be avoided.

Series with general terms which are the product of two quantities can be dealt with by two more useful tests, *Dirichlet's test*, Theorem 6.55, and *Abel's test*, Theorem 6.57. Both are applications of the formula of *summation by parts*, Proposition 6.51.

Product of series

Let $\mathbf{a} := \{a_n\}$ and $\mathbf{b} := \{b_n\}$, $n \geq 0$, be two sequences. The sequence, denoted by $\{(\mathbf{a} * \mathbf{b})_n\}$,

$$(\mathbf{a} * \mathbf{b})_n := \sum_{i+j=n} a_i b_j = \sum_{j=0}^n a_j b_{n-j}$$

is called the *product of convolution*, or *Cauchy product*, of \mathbf{a} and \mathbf{b} .

- **CAUCHY'S THEOREM.** If $\sum_{j=0}^{\infty} a_j$ and $\sum_{j=0}^{\infty} b_j$ converge absolutely, then

$$\sum_{j=0}^{\infty} (\mathbf{a} * \mathbf{b})_j$$

converges absolutely and

$$\sum_{j=0}^{\infty} (\mathbf{a} * \mathbf{b})_j = \left(\sum_{j=0}^{\infty} a_j \right) \left(\sum_{j=0}^{\infty} b_j \right).$$

This extends the usual formula for the product of two polynomials to a couple of absolutely convergent series.

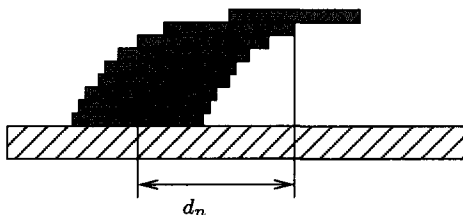


Figure 6.21. The problem of the pile of coins.

6.9 Exercises

6.68 ¶. Compute the sums of the telescoping series

$$\sum_{j=1}^{\infty} \frac{\sqrt{j+1} - \sqrt{j}}{\sqrt{j^2 + j}}, \quad \sum_{j=1}^{\infty} \frac{j+1/2}{j^2(j+1)^2}, \quad \sum_{j=1}^{\infty} \frac{1}{j(j+1)(j+2)}, \quad \sum_{j=2}^{\infty} \frac{1}{j^2-1}.$$

6.69 ¶. A ball falls from height h onto a rigid surface. It rebounds infinitely many times, each time reaching 75% of the height of the previous rebound. Compute the time needed in order that the ball be at rest.

6.70 ¶ von Koch's curve. Starting from an equilateral triangle, erect an equilateral triangle on the middle third of its sides. Iterate the process on each of the sides of the polygonal figure obtained this way. The limit closed curve defined this way is called *von Koch's curve*. Show that the resulting area is finite and compute it. Show that, instead, von Koch's curve has infinite length.

6.71 ¶ Cantor set. A unit square is divided in 9 squares of side $1/3$, the central one is colored black and the remaining 8 are divided each in 9 squares of side $1/9$, and each of the central squares is coloured black. By induction we now iterate the process infinitely many times. Compute the area of the black region. The complement in the unit square of the black region is known as a *Cantor set*.

6.72 ¶. Estimate the error we make replacing $\sum_{j=1}^{\infty} 1/j^2$ with one of its partial sums.

6.73 ¶. A slow caterpillar is crawling on a rubber band at the speed of 1cm per minute, but a malevolent elf lengthens the band by 1m per minute. Will the caterpillar ever be able to reach the end of the rubber band?

6.74 ¶. Make a pile of n coins of diameter 1. Dislodge them all in the same direction as much as possible and keep them in equilibrium, as shown in Figure 6.21. What is the horizontal distance $\{d_n\}$ between the centers of the first and the last coin?

6.75 ¶. Show the following

Proposition (Asymptotic root test). Let $\{a_n\}$ be a sequence of nonnegative numbers. If

$$\limsup_{n \rightarrow \infty} \sqrt[n]{a_n} < 1,$$

then $\sum_{n=0}^{\infty} a_n$ converges. If

$$\limsup_{n \rightarrow \infty} \sqrt[n]{a_n} > 1,$$

then $\sum_{n=0}^{\infty} a_n$ diverges.

Analytical formulas for π

- VIÈTE. $\frac{2}{\pi} = \sqrt{\frac{1}{2}} \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}} \sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2} + \frac{1}{2}\sqrt{\frac{1}{2}}}} \dots$
- WALLIS. $\frac{\pi}{2} = \prod_{n=1}^{\infty} \frac{2n}{2n-1} \frac{2n}{2n+1}$
- GREGORY, LEIBNIZ. $\frac{\pi}{4} = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1}$
- NEWTON. $\frac{\pi}{6} = \sum_{n=0}^{\infty} \frac{1}{(2n+1)2^{2n+1}}$
- $\frac{\pi}{2\sqrt{3}} = \sum_{n=0}^{\infty} (-1)^n \frac{1}{3^n(2n+1)}$
- $\frac{\pi}{4} = \sum_{n=0}^{\infty} (-1)^n \frac{1}{2n+1} \left(\frac{4}{5^{n+1}} \right) - \frac{1}{(239)^{n+1}}$
- AREA OF THE UNIT CIRCLE. $\frac{\pi}{2} = \int_{-1}^1 \sqrt{1-x^2} dx$
- HALF THE LENGTH OF THE CIRCLE. $\pi = \int_{-1}^1 \frac{1}{\sqrt{1-x^2}} dx$

The number e

- We defined in [GM1] the Euler number e as the unique real number such that $\int_0^e \frac{1}{t} dt = 1$. We know, see e.g., [GM1], that

$$D(e^x) = e^x, \quad \text{equivalently} \quad \lim_{x \rightarrow 0} \frac{e^x - 1}{x} = 1,$$

hence,

$$e^x = \lim_{n \rightarrow \infty} \left(1 + \frac{x}{n} \right)^n.$$

- We have

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \quad \forall x \in \mathbb{R} \quad \text{with} \quad \left| e^x - \sum_{k=0}^n \frac{x^k}{k!} \right| \leq \max(e^x, 1) \frac{|x|^{n+1}}{(n+1)!},$$

$$e = \sum_{n=0}^{\infty} \frac{1}{n!} \quad \text{with} \quad 0 < e - \sum_{k=0}^n \frac{1}{k!} < \frac{1}{n n!}.$$

- e is irrational.

Viète and Euler formulas for $\sin x$

$$\sin x = x \prod_{k=1}^{\infty} \cos\left(\frac{x}{2^k}\right), \quad \sin x = x \prod_{j=1}^{\infty} \left(1 - \frac{x^2}{j^2 \pi^2}\right).$$

Figure 6.22. Analytical formulas for π , e , and the Viète and Euler formulas for $\sin x$.

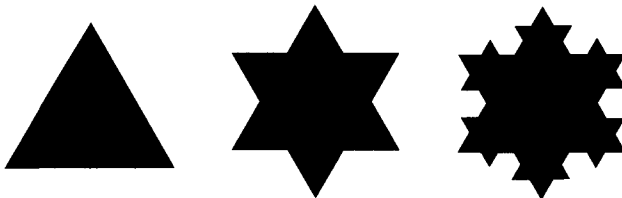


Figure 6.23. The first steps in the construction of the von Koch curve.

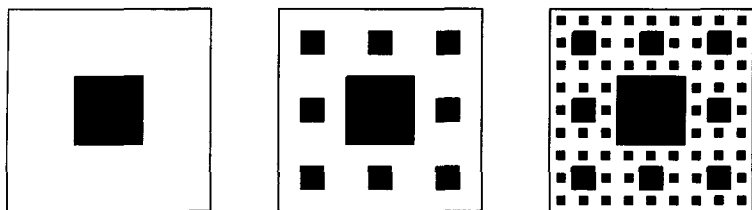


Figure 6.24. The first steps in the construction of the Cantor set in Exercise 6.71.

6.76 ¶. Show the following

Proposition (Asymptotic ratio test). Let $\{a_n\}$ be a sequence of positive numbers. If

$$\limsup_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} < 1,$$

then $\sum_{n=0}^{\infty} a_n$ converges. If

$$\limsup_{n \rightarrow \infty} \frac{a_{n+1}}{a_n} > 1,$$

then $\sum_{n=0}^{\infty} a_n$ diverges.

6.77 ¶. Show that

- (i) if $n^2(a_{n+1} - a_n) \rightarrow 0$, then $\{a_n\}$ converges,
- (ii) if $\sum_{n=1}^{\infty} a_n^2$ converges, then $\sum_{n=1}^{\infty} a_n/n$ converges, too.

6.78 ¶. Show that $\sum_{n=1}^{\infty} \frac{z^n}{n}$ converges iff $z \in \mathbb{C}$, $|z| \leq 1$ and $z \neq 1$, and that $\sum_{n=1}^{\infty} (-1)^n \frac{z^{2n}}{n}$ converges if and only if $z \in \mathbb{C}$, $|z| \leq 1$ and $z \neq \pm i$.

6.79 ¶ **Pringsheim's theorem.** Let $\{a_n\}$ be a decreasing sequence such that $\sum_{n=0}^{\infty} a_n$ converges. Show that $na_n \rightarrow 0$ as $n \rightarrow \infty$.

6.80 ¶ **Kronecker's lemma.** If $\sum_{n=1}^{\infty} a_n/n$ converges, then $\frac{1}{N} \sum_{n=1}^N a_n \rightarrow 0$ as $n \rightarrow \infty$.

6.81 ¶. Suppose $a_n \rightarrow A$ and $b_n \rightarrow B$. Show that $\frac{1}{n}(a * b)_n \rightarrow AB$. [Hint: Write $a_n := A + \epsilon_n$.]

6.82 ¶. Let $a_n \geq 0$ and $\lambda > 1$. Show that

$$\sum_{n=1}^{\infty} \frac{a_n}{\left(\sum_{k=1}^n a_k\right)^{\lambda}}$$

converges; estimate its sum. [Hint: Write the terms in function of $S_n := \sum_{k=1}^n a_k$.]

6.83 ¶. Let $\{a_n\}$ and $\{b_n\}$ be two monotone and infinitesimal sequences. Show that $\sum_{n=1}^{\infty} a_n \sin nt$, $t \in \mathbb{R}$, and $\sum_{n=1}^{\infty} b_n \cos nt$, $t \neq 0$, converge. [Hint: Use Dirichlet's test.]

6.84 ¶. Is there a sequence $\{a_n\}$ such that $\sum_{n=1}^{\infty} \left(a_n + \frac{\epsilon^{a_n}}{a_n}\right)$ converges?

6.85 ¶. Let $\alpha > 1$. Find a lower bound for $s_k := \sum_{n=1}^k \frac{1}{n^{\alpha}}$.

6.86 ¶ Eisenstein series. Study the convergence of the complex series $\sum_{n=1}^{\infty} \left(\frac{1}{z+n} - \frac{1}{n} \right)$, $\sum_{n=1}^{\infty} \left(\frac{1}{z-n} + \frac{1}{n} \right)$, $\sum_{n=0}^{\infty} \frac{1}{(z+n)^k}$, $\sum_{n=0}^{\infty} \frac{1}{(z-n)^k}$, where $k \geq 2$. [Hint: Observe that, for $r > 0$, $|z \pm n|^k \geq (n-r)^k$ for $k \geq 1$, $n \in \mathbb{N}$ and $|z| < r < n$.]

6.87 ¶. Study the convergence of some of the following series, possibly dependent on the real parameter x or on the complex parameter z :

$$\begin{array}{ll} ne^{-n}, & n^{-n}, \\ \frac{n!}{(2n)!}, & \frac{n^2 \log(1+n^2)}{\sqrt{n!}}, \\ \frac{(2n)!}{(3n)! - (2n)!}, & \frac{2^n n!}{(2n)!}, \\ \frac{\cos \pi n}{n \log(1+n)}, & \frac{\sin n!}{n(n+1)}, \\ \arctan 2^{-n}, & \frac{(-1)^k k}{(k+1)(k+2)}, \\ \log^2(1+1/n), & \log(1+1/n), \\ \frac{n}{n \arctan n + 1}, & (-1)^n \sin(1/n), \end{array}$$

6.88 ¶. Study the convergence of some of the following series, possibly dependent on the real parameter x or on the complex parameter z :

$$\begin{array}{ll} n^{\pi \sin(1/n) - 2\pi}, & \left(1 - \frac{1}{\sqrt{n}}\right)^{\sqrt{n^3}}, \\ \frac{(-1)^n}{n - \log n}, & (\log n)^{-\log n}, \\ \left(\frac{\pi}{2} - \arctan n\right), & \log\left(1 + \frac{(-1)^n}{n}\right), \\ \log\left(1 + \frac{(-1)^n}{\sqrt{n}}\right), & \sin\left(\frac{(-1)^n}{\sqrt{n}}\right), \\ (1+x)^{n(1+1/n)}, \ x \in]-1, 0[, & (-2)^n e^{-nx}, \\ a^{n^2} z^n, & \frac{z^n}{n^n}, \\ \frac{n!}{n^n} z^n, & (1 - \sin(1/n))^{1/n}, \\ (-1)^n \log\left(1 + \frac{(-1)^n}{n}\right), & \log^2 \frac{n}{n+1}, \\ \left(1 + \frac{1}{n^2}\right)^{n^2} - e, & \left(\frac{\cos(2/n)}{\cos(1/n)}\right)^{n^3}, \\ \left(\sin \frac{1}{n}\right)^{2 \cos(1/n)}, & \frac{\sqrt[3]{n} - 1}{\log n}, \\ \frac{\pi}{4} - \arctan \cos \frac{1}{n}, & \left(x \arctan n + \frac{1}{n^2}\right)^{n+1}, \\ \frac{e^{x/n} - 1}{nx}, & \int_{n^2}^{n^3} \sin^2 \frac{1}{x} dx, \\ \int_0^{1/n} \left(1 - \frac{\sin x}{x}\right) dx, & \int_0^{\pi/2} (\sin x)^n dx, \end{array}$$

$$\begin{aligned}
 &(-1)^n \int_n^{n+1} t^2 e^{-t^2} dt, & n \int_n^{n+1} e^{-x} \sin x dx, \\
 &\sqrt{n} \int_n^{n+1} \frac{\sin x}{x^2} dx.
 \end{aligned}$$

6.89 ¶. For some of the previous series, estimate their sum or the order of magnitude of their partial sums.

6.90 ¶¶ Raabe's test. Let $\sum_{n=1}^{\infty} a_n$ be a series of positive terms. Show that, if there exists $K > 1$ such that

$$n \left(\frac{a_n}{a_{n+1}} - 1 \right) \geq K \quad \forall n,$$

then $\sum_{n=1}^{\infty} a_n$ converges, while, if

$$n \left(\frac{a_n}{a_{n+1}} - 1 \right) \leq 1 \quad \forall n,$$

then $\sum_{n=1}^{\infty} a_n$ diverges. [Hint: In the first case show that $a_{n+1} \leq \frac{1}{4}(na_n - (n+1)a_{n+1})$; in the second show that $a_{n+1} \geq a_1/(n+1)$.]

6.91 ¶¶ Gauss's test. Let $\sum_{n=1}^{\infty} a_n$ be a series of positive terms. Suppose that

$$\frac{a_n}{a_{n+1}} = 1 - \frac{K}{n} + \frac{\theta_n}{n^{1+p}}$$

where $p > 0$ and $\{\theta_n\}$ is a bounded sequence. Then $\sum_{n=1}^{\infty} a_n$ converges if $K > 1$ and diverges if $K \leq 1$.

6.92 ¶¶. Discuss the convergence of the following series:

$$\begin{aligned}
 &\sum_{n=1}^{\infty} \frac{n!}{(\alpha+1)(\alpha+2)\cdots(\alpha+n)}, & \alpha \text{ positive integer,} \\
 &\sum_{n=1}^{\infty} \frac{1 \cdot 4 \cdot 7 \cdots (3n-2)}{3 \cdot 6 \cdot 9 \cdots 3n}, \\
 &\sum_{n=1}^{\infty} \left(\frac{1 \cdot 4 \cdot 7 \cdots (3n-2)}{3 \cdot 6 \cdot 9 \cdots 3n} \right)^2.
 \end{aligned}$$

6.93 ¶. Let $u_n = v_n := (-1)^n / \sqrt{n+1}$. The series $\sum_{n=0}^{\infty} u_n$ and $\sum_{n=0}^{\infty} v_n$ converge, though not absolutely. Show that the product series $\sum_{n=0}^{\infty} w_n$,

$$w_n := (-1)^n \sum_{k=0}^n \frac{1}{(k+1)(n+1-k)},$$

does not converge. [Hint: Show that $|w_n| \geq \frac{2n+2}{n+2}$.]

6.94 ¶. Let $\lambda_k \geq 0$ and $\sum_{k=0}^{\infty} \lambda_k < \infty$. Find a sequence of positive numbers $\{\alpha_k\}$, $\alpha_k \rightarrow \infty$, such that $\sum_{k=0}^{\infty} \alpha_k \lambda_k < \infty$.

7. Power Series

The manipulation of series reaches such levels of subtlety that is hard to imagine even nowadays in the works of Jacob Bernoulli (1654–1705), Johann Bernoulli (1667–1748) and Leonhard Euler (1707–1783), and particularly in the *Ars conjectandi* by Jacob Bernoulli in *Introductio in analysin infinitorum* and in *Institutiones calculi* by Euler. For instance in *Ars conjectandi* Jacob Bernoulli introduced the so-called *Bernoulli's numbers*, see Section 7.3 below, which may be defined by

$$\frac{x}{e^x - 1} = \sum_{n=0}^{\infty} B_n \frac{x^n}{n!};$$

Euler proved that

$$\sum_{n=1}^{\infty} \frac{1}{n^{2k}} = \frac{(2\pi)^{2k}}{2(2k)!} |B_{2k}|$$

that in particular yields

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}, \quad \sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90}.$$

Euler also proved that

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n^{2k}} = \frac{\pi^{2k}(2^{2k} - 1)}{(2k)!} |B_{2k}|, \quad \sum_{n=1}^{\infty} \frac{1}{(2n+1)^{2k}} = \frac{\pi^{2k}2^{2k-1}}{2(2k)!} |B_{2k}|$$

that yields for example

$$1 - \frac{1}{2^2} + \frac{1}{3^2} - \frac{1}{4^2} + \cdots = \sum_{n=1}^{\infty} \frac{(-1)^{n-1}}{n^2} = \frac{\pi^2}{12}$$

$$1 + \frac{1}{3^2} + \frac{1}{5^2} + \cdots = \sum_{n=1}^{\infty} \frac{1}{(2n+1)^2} = \frac{\pi^2}{8}.$$

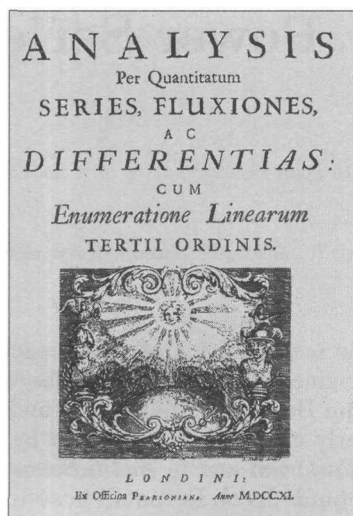
Still Euler, introducing the so-called *Euler's numbers* as²

¹ We have $B_0 = 1$, $B_1 = -1/2$, $B_2 = 1/6$, $B_{2n+1} = 0 \forall n \geq 1$.

² $E_{2n-1} = 0 \forall n \geq 1$ e $E_{2n} = 4^{2n+1}(B_n - 1/4)^{2n+1}/(2n+1)$.



Figure 7.1. Leonhard Euler (1707–1783) and the frontispiece of the *Analysis* by Sir Isaac Newton (1643–1727).



$$\frac{1}{\cosh x} = \sum_{n=0}^{\infty} E_n \frac{x^n}{n!}$$

found that

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)^{2k+1}} = \frac{\pi^{2k+1}}{2^{2k+2}(2k)!} |E_{2k}|$$

from which

$$1 - \frac{1}{3^3} + \frac{1}{5^3} - \cdots = \sum_{n=1}^{\infty} \frac{(-1)^n}{(2n+1)^3} = \frac{\pi^3}{32}.$$

Euler, as many others, treated also *divergent series* finding their asymptotic behaviour³.

In the eighteenth century the use of series was however quite formal, not much attention was given to their convergence, though it was not totally ignored. It is only in the beginning of the nineteenth century that series were treated correctly with Joseph Fourier (1768–1830), Carl Friedrich Gauss (1777–1855), Bernhard Bolzano (1781–1848), Niels Henrik Abel (1802–1829).

A satisfactory definition of convergence appeared in the *Théorie analytique del la chaleur*, but the first rigorous definition is due to Gauss in the

³ Divergent series play a fundamental role in the study of differential equations, as, for instance, in the study of the vibrations of membranes with Wilhelm Bessel (1784–1846), Enrico Betti (1823–1892), Thomas Jan Stieltjes (1856–1894) and Ernesto Cesàro (1859–1906) among others, and starting from J. Henri Poincaré (1854–1912), in the theory of perturbations of integrable differential systems.

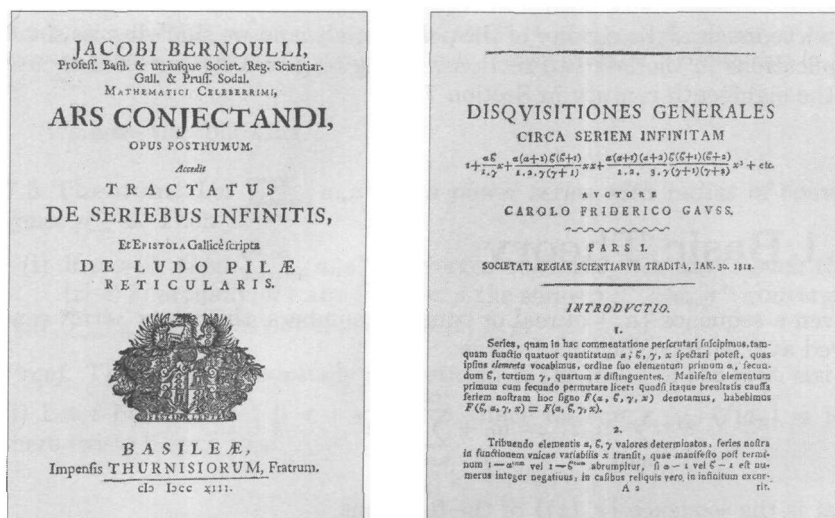


Figure 7.2. The frontispiece of *Ars conjectandi* by Jacob Bernoulli (1654–1705) and the first page of the *Disquisitiones* by Carl Friedrich Gauss (1777–1855).

paper *Disquisitiones generales* in 1812, in which Gauss studies the hypergeometric series already discussed by Euler. Finally, it was with Bernhard Bolzano (1781–1848) and, especially, with Augustin-Louis Cauchy (1789–1857) in his *Cours d'Analyse* that the theory of series found its firm basis. With Karl Weierstrass (1815–1897) the theory of complex power series identifies with the theory of *complex analytic functions*, a theory which is extremely relevant in physics and engineering, as well as in algebraic geometry and analytic number theory. In connection with number theory we should at least mention *Dirichlet series*

$$\sum_{n=1}^{\infty} \frac{a_n}{n^z}$$

the simplest of which is

$$\zeta(z) := \sum_{n=1}^{\infty} \frac{1}{n^z}$$

that defines for $\Re(z) > 1$ *Riemann's zeta function*, of tremendous relevance in studying the distribution of prime numbers as

$$\zeta(z) = \prod_{p \text{ prime}} \frac{1}{1 - p^{-z}},$$

as well as in studying the distribution of the eigenvalues of differential operators.

Of course our goal in this chapter is a great deal more modest. We shall develop the basic theory in the first section, where we see how power series

preserve much of the rigidity of the polynomials, and we shall discuss some applications in the last two sections, trying to give a flavour of their use in the eighteenth century in Section 7.4.

7.1 Basic Theory

Given a sequence $\{a_n\}$ of real or complex numbers, the *power series* centered at 0 with coefficients $\{a_n\}$ is

$$\sum_{n=0}^{\infty} a_n z^n := a_0 + \sum_{n=1}^{\infty} a_n z^n, \quad z \in \mathbb{C},$$

that is the sequence $\{s_n(z)\}$ of the functions

$$s_n(z) = \sum_{k=0}^n a_k z^k := a_0 + \sum_{k=1}^n a_k z^k, \quad z \in \mathbb{C}.$$

We might as well consider the power series centered at z_0 with coefficients a_n ,

$$\sum_{n=0}^{\infty} a_n (z - z_0)^n,$$

but of course the two series are related by a simple change of variables. Also, If we restrict ourselves to the real axis x , we may as well consider the real power series $\sum_{n=0}^{\infty} a_n x^n$, $x \in \mathbb{R}$.

Clearly the series $\sum_{n=0}^{\infty} a_n z^n$ converges or diverges depending on the choice of z . $\sum_{n=0}^{\infty} a_n z^n$ obviously converges at zero with sum a_0 . What is special of power series is that to each of them is associated a *disc of convergence* such that the series converges if z is in the interior of the disc (provided the radius of that disc is positive) and diverges if z is in the exterior of the disc.

7.1.1 Circle of convergence

7.1 Definition. Let $\sum_{n=0}^{\infty} a_n z^n$ be a power series. The number $\rho \in \overline{\mathbb{R}}_+$ defined by

$$\frac{1}{\rho} := \limsup_{n \rightarrow \infty} \sqrt[n]{|a_n|},$$

is called the radius of convergence of $\sum_{n=0}^{\infty} a_n z^n$. We use the conventions $1/0^+ = +\infty$ and $1/+\infty = 0$.

7.2 ¶. A series $\sum_{n=0}^{\infty} a_n z^n$ has a positive radius of convergence if and only if $\{|a_n|\}$ grows at most exponentially, e.g., if and only if there exists $M > 0$ such that $|a_n| \leq M^n \forall n$.

We have the following.

7.3 Theorem. Let $\sum_{n=0}^{\infty} a_n z^n$ be a power series with radius of convergence $\rho \geq 0$. Then

- (i) if $\rho > 0$, then $\sum_{n=0}^{\infty} a_n z^n$ converges absolutely for any z such that $|z| < \rho$; actually, for any $0 < r < \rho$ the series $\sum_{n=0}^{\infty} |a_n| r^n$ converges,
- (ii) $\sum_{n=0}^{\infty} a_n z^n$ does not converge if $|z| > \rho$.

Proof. The proof is essentially a repetition of the proof of the root test.

- (i) Let t be such that $0 < r < t < \rho$. Since $\limsup_{n \rightarrow +\infty} \sqrt[n]{|a_n|} = 1/\rho$ there exists \bar{n} such that

$$\sqrt[n]{|a_n|} < \frac{1}{t}$$

for all $n \geq \bar{n}$, hence

$$\sqrt[n]{|a_n|} r^n \leq \frac{r}{t} < 1$$

from which

$$|a_n| r^n \leq \left(\frac{r}{t}\right)^n \quad \text{for all } n \geq \bar{n}.$$

A comparison with the geometric series yields the second claim and the estimate

$$\sum_{n=p}^{\infty} |a_n| r^n \leq \sum_{n=p}^{\infty} h^n = \frac{\left(\frac{r}{t}\right)^p}{1 - \frac{r}{t}} \quad \forall p \geq \bar{n}, \quad (7.1)$$

and therefore (i).

- (ii) If $|z| > \rho$, then $\limsup_{n \rightarrow \infty} \sqrt[n]{|a_n| |z|^n} = |z|/\rho > 1$. From the characteristic property of \limsup , if h is a number between 1 and $|z|/\rho$, we can find a subsequence $\{a_{k_n}\}$ of $\{a_n\}$ such that $\sqrt[k_n]{|a_{k_n}|} |z| \geq h \forall n$, i.e.,

$$|a_{k_n}| |z|^{k_n} \geq h^{k_n} \quad \forall n.$$

It follows that $|a_{k_n}| |z|^{k_n} \rightarrow \infty$. In particular, $a_n z^n$ does not converge to zero, consequently by the following.

Proposition 6.7 the series does not converge at z . □

7.4 Remark. Theorem 7.3 implies the following characterizations for the radius of convergence ρ of $\sum_{n=0}^{\infty} a_n z^n$:

$$\begin{aligned} \rho &:= \sup \left\{ |z| \mid \sum_{n=0}^{\infty} a_n z^n \text{ converges absolutely} \right\} \\ &= \sup \left\{ |z| \mid \sum_{n=0}^{\infty} a_n z^n \text{ converges} \right\}. \end{aligned}$$

a. The disc and the domain of convergence

Theorem 7.3 implies that the *domain of convergence* of a power series, that is the set Δ of points in which it converges, is its disc of convergence $\{z \mid |z| < \rho\}$, (the open interval $] -\rho, \rho[$ for real series) union possibly one or more points on the circle $|z| = \rho$ (one or both of $-\rho, \rho$ for real series)

$$\{z \mid |z| < \rho\} \subset \Delta \subset \{z \mid |z| \leq 1\}.$$

Let us see a few examples.

7.5 Example. We already saw several times that the geometric series $\sum_{n=0}^{\infty} x^n$ converges if and only if $|x| < 1$ and that

$$\sum_{n=1}^{\infty} x^n = \frac{1}{1-x}, \quad |x| < 1.$$

The domain of convergence of the geometrical series is then the open interval $] -1, 1[$.

Similarly we proved in Example 6.40 that the complex geometric series $\sum_{n=0}^{\infty} z^n$ converges if and only if $|z| < 1$. This time the domain of convergence is the interior of the disc of convergence, $\{z \mid |z| < 1\}$.

7.6 Example. The power series $\sum_{n=0}^{\infty} x^n/n^2$ converges absolutely, hence converges, for all x such that $|x| \leq 1$, since in this case $|x|^n/n^2 \leq 1/n^2$ for all n and $\sum_{n=1}^{\infty} 1/n^2$ converges. $\sum_{n=0}^{\infty} x^n/n^2$ does not converge if $|x| > 1$ by Proposition 6.7 since in this case $|x|^n/n^2 \rightarrow \infty$. Thus $\sum_{n=1}^{\infty} x^n/n^2$ converges (actually converges absolutely) if and only if $|x| \leq 1$. This time the domain of convergence is the closed interval $[-1, 1]$.

With exactly the same computations, one can show that the complex series $\sum_{n=1}^{\infty} z^n/n^2$ converges absolutely if and only if $|z| \leq 1$ and does not converge if $|z| > 1$. Thus the domain of convergence is the disc of convergence union its boundary.

7.7 Example. We saw in Example 6.13 that $\sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1}$ converges if and only if $-1 < x \leq 1$ with sum $\log(1+x)$. Thus $\sum_{n=0}^{\infty} (-1)^n x^{n+1}/(n+1)$ has radius of convergence $\rho = 1$, and its domain of convergence is the half-open interval $] -1, 1]$.

The complex series $\sum_{n=0}^{\infty} (-1)^n \frac{z^{n+1}}{n+1}$ has the same radius of convergence $\rho = 1$. Thus it converges absolutely if $|z| < 1$ and does not converge if $|z| > 1$. It can be proved as a trivial application of the Dirichlet test, Theorem 7.30, that $\sum_{n=1}^{\infty} (-1)^n \frac{z^{n+1}}{n+1}$ converges at any z such that $|z| = 1$ except $z = -1$. Therefore this time the domain of convergence is $\{z \mid |z| \leq 1, z \neq -1\}$.

7.8 Example. The power series $\sum_{n=1}^{\infty} \frac{2^n}{n^2} z^{2n}$ has coefficients

$$a_n = \begin{cases} 0 & \text{if } n \text{ is odd,} \\ 2^p/p^2 & \text{if } n = 2p, \end{cases}$$

and radius of convergence $1/\sqrt{2}$ since

$$\frac{1}{\rho} = \limsup_{n \rightarrow \infty} \sqrt[n]{a_n} = \lim_{p \rightarrow \infty} \sqrt[2p]{\frac{2^p}{p^2}} = \sqrt{2}.$$

We may also regard $\sum_{j=1}^{\infty} (2^n/n^2) z^{2n}$ as a power series in $2z^2$. In other words, by changing variable, $w := 2z^2$, we get the power series in Example 7.6, $\sum_{n=1}^{\infty} w^n/n^2$, that converges absolutely for all w such that $|w| = 2|z|^2 \leq 1$, concluding that $\sum_{n=1}^{\infty} (2^n/n^2) z^{2n}$ converges absolutely for $|z| \leq 1/\sqrt{2}$ and does not converge for $|z| > 1/\sqrt{2}$.

7.1.2 Continuity of the sum

Let $\sum_{n=0}^{\infty} a_n z^n$ be a power series with radius of convergence $\rho > 0$. For any n , $S_n(z) := \sum_{j=0}^n a_j z^j$ is a polynomial, it is therefore natural to ask how many properties of the functions $S_n(z)$ are preserved when passing to the limit. In particular, is the sum of the series $S(z)$ continuous in its domain of definition? It is hard to resist writing

$$\begin{aligned} \lim_{z \rightarrow z_0} S(z) &= \lim_{z \rightarrow z_0} \lim_{n \rightarrow \infty} S_n(z) \\ &= \lim_{n \rightarrow \infty} \lim_{z \rightarrow z_0} S_n(z) = \lim_{n \rightarrow \infty} S_n(z_0) = S(z_0). \end{aligned} \quad (7.2)$$

However, it is a fact that the change in the order is not allowed, i.e., the equality

$$\lim_{z \rightarrow z_0} \lim_{n \rightarrow \infty} S_n(z) = \lim_{n \rightarrow \infty} \lim_{z \rightarrow z_0} S_n(z)$$

is in general false. For instance, if $S_n(x) = x^n$, $x \in [0, 1]$, then

$$\lim_{x \rightarrow 1^-} \lim_{n \rightarrow \infty} x^n = \lim_{x \rightarrow 1^-} 0 = 0 \quad \text{and} \quad \lim_{n \rightarrow \infty} \lim_{x \rightarrow 1^-} x^n = \lim_{n \rightarrow \infty} 1 = 1. \quad (7.3)$$

The computation in (7.2) is therefore unjustified.

a. Uniform Convergence

The exchange of limits in (7.2) turns out to be correct in case we have a *uniform* estimate (in z) for the error when replacing $S(z)$ by $S_n(z)$. For the relevance of this notion it is worth giving the following

7.9 Definition. Let $\{S_n\}$ be a sequence of functions $S_n : A \rightarrow \mathbb{C}$ defined on a subset A of \mathbb{R}^2 , and let $S : A \rightarrow \mathbb{C}$. We say that $\{S_n\}$ converges uniformly to S in A if

$$\forall \epsilon > 0 \exists \bar{n} \text{ such that } |S_n(z) - S(z)| < \epsilon \quad \forall n \geq \bar{n} \text{ and } \forall z \in A.$$

We say that a series of functions $\sum_{n=1}^{\infty} f_n(z)$ converges uniformly in A to $f : A \rightarrow \mathbb{C}$, if the sequence of its partial sums converges uniformly in A .

7.10 Remark. In comparison with the *pointwise convergence* in A , that is $S_n(z) \rightarrow S(z) \quad \forall z \in A$, the uniform convergence says (requires) that the index \bar{n} , for which the error $|S_n(z) - S(z)|$ is smaller than ϵ for all $n \geq \bar{n}$, does not depend on the particular point $z \in A$.

Notice that the definition does not allow us to produce a uniform limit, but it only allows us to verify whether a function $S : A \rightarrow \mathbb{R}$ is or is not the uniform limit of the sequence $\{S_n\}$. According to Definition 7.9, computing uniform limits is a two-step procedure:

- first, guess a possible limit function $S : A \rightarrow \mathbb{R}$,
- second, prove that S is in fact the uniform limit of $\{S_n\}$ in A .

By considering the maximal error between $S_n(z)$ and $S(z)$ when z varies in A ,

$$\|S_n - S\|_{\infty, A} := \sup_{z \in A} |S_n(z) - S(z)|,$$

we can say, by comparing the explicit definitions, that $S_n \rightarrow S$ uniformly if and only if the numerical sequence $\{\|S_n - S\|_{\infty, A}\} \rightarrow 0$ as $n \rightarrow \infty$. Also, from

$$|S_n(z) - S(z)| \leq \|S_n - S\|_{\infty, A} \quad \forall z \in A,$$

uniform convergence of $\{S_n\}$ to S in A implies pointwise convergence at each point $z \in A$. As consequence the pointwise limit is the only possible candidate to be the uniform limit.

The converse of the last claim is false in general, since there exist sequences of functions that converge pointwisely but not uniformly, as the following example shows.

7.11 Example. Consider a function $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ that is bounded with $\|\varphi\|_{\infty, \mathbb{R}} = M > 0$ and such that $\varphi(x) \rightarrow 0$ as $x \rightarrow -\infty$. Define $S_n(x) := \varphi(x - n)$, $x \in \mathbb{R}$. We see that for any fixed x , $S_n(x) \rightarrow 0$, hence $\{S_n\}$ converges pointwisely to 0, but $\{S_n\}$ does not converge uniformly to zero since $\|\Sigma_n\|_{\infty, \mathbb{R}} = \|\varphi\|_{\infty, \mathbb{R}} = M > 0$.

7.12 ¶. Show that the sequence $\{\varphi_n\}$ of functions $\varphi_n : [0, 1] \rightarrow \mathbb{R}$ given by $\varphi(x) := \frac{x}{n}e^{-x/n}$ converges to zero in $[0, 1]$ but not uniformly.

b. Continuity of uniform limits

7.13 Theorem. Let $\{S_n\}$ be a sequence of continuous functions $S_n : A \rightarrow \mathbb{C}$ on A and suppose that $\{S_n\}$ converges uniformly in A to $S : A \rightarrow \mathbb{C}$. Then S is continuous on A .

Proof. Let $z_0 \in A$ and $\epsilon > 0$. Since $S_n(z) \rightarrow S(z)$ uniformly, there exists n such that $|S_n(z) - S(z)| < \epsilon$ for all $z \in A$. Consequently

$$\begin{aligned} |S(z) - S(z_0)| &= |S_n(z) - S_n(z_0) + S(z) - S_n(z) + S_n(z_0) - S(z_0)| \\ &\leq |S_n(z) - S_n(z_0)| + |S(z) - S_n(z)| + |S(z_0) - S_n(z_0)| \\ &\leq |S_n(z) - S_n(z_0)| + 2\epsilon. \end{aligned}$$

Since S_n is continuous, we also find $\delta > 0$ such that $|S_n(z) - S_n(z_0)| < \epsilon$ whenever $z \in A$ and $|z - z_0| < \delta$. In conclusion we infer that $|S(z) - S(z_0)| < 3\epsilon$ for all $z \in A$ with $|z - z_0| < \delta$. \square

Notice that the assumption of uniform convergence in Theorem 7.13 cannot be dropped. For instance, the sequence $\{x^n\}$, $x \in [0, 1]$, converges pointwisely to the discontinuous function

$$S(x) = \begin{cases} 0 & \text{if } 0 \leq x < 1, \\ 1 & \text{if } x = 1. \end{cases}$$

c. Uniform convergence of power series

Going back to power series, we have the following.

7.14 Theorem. Let $\sum_{n=0}^{\infty} a_n z^n$ be a power series that converges absolutely at z_0 . Then $\sum_{n=0}^{\infty} a_n z^n$ converges uniformly in $\{z \mid |z| \leq |z_0|\}$. In particular, if ρ is the radius of convergence of $\sum_{n=0}^{\infty} a_n z^n$ and $\rho > 0$, then $\sum_{n=0}^{\infty} a_n z^n$ converges uniformly in $\{z \mid |z| \leq r\}$ for all $r < \rho$.

Proof. For $|z| \leq |z_0|$ and $n \geq 1$ we have by the triangular inequality

$$|S(z) - S_n(z)| = \left| \sum_{j=n+1}^{\infty} a_j z^j \right| \leq \sum_{j=n+1}^{\infty} |a_j| |z|^j \leq \sum_{j=n+1}^{\infty} |a_j| |z_0|^j, \quad (7.4)$$

hence

$$\sup_{|z| \leq |z_0|} |S(z) - S_n(z)| \leq \sum_{j=n+1}^{\infty} |a_j| |z_0|^j \rightarrow 0, \quad n \rightarrow \infty.$$

The second part of the claim follows from Theorem 7.3. \square

Theorems 7.14 and 7.13 then yield at once

7.15 Corollary. Let $S(z) = \sum_{n=0}^{\infty} a_n z^n$ be the sum of a power series with a positive radius of convergence $\rho > 0$. Then $S(z)$ is continuous on $\{z \mid |z| < \rho\}$.

Proof. Let z_0 be inside the disc of convergence, $|z_0| < \rho$, and let s be such that $|z_0| < s < \rho$. By Theorems 7.14 and 7.13 the restriction of $S(z)$ to $|z| \leq s$ is continuous. Consequently S is continuous at z_0 , since $|z_0| < s$. \square

7.16 ¶. Show directly, that is, by using the definition of continuity, that the sum of a power series is continuous on $\{z \mid |z| < \rho\}$. [Hint: Let z_0 be such that $|z_0| < \rho$. Observe that for $\sigma < \rho - |z_0|$ and $|z - z_0| < \sigma$ one has

$$\begin{aligned} \left| \sum_{n=0}^{\infty} a_n z^n - \sum_{n=0}^{\infty} a_n z_0^n \right| &= \left| \sum_{n=1}^{\infty} a_n (z^n - z_0^n) \right| \\ &\leq \left| \sum_{n=1}^{\infty} \left\{ a_n (z - z_0) (z^{n-1} + z^{n-2} z_0 + \cdots + z z_0^{n-2} + z_0^{n-1}) \right\} \right| \\ &\leq \sum_{n=1}^{\infty} |a_n| |z - z_0| n (|z_0| + \sigma)^{n-1} = |z - z_0| \sum_{n=1}^{\infty} n |a_n| (|z_0| + \sigma)^{n-1}. \end{aligned}$$

7.1.3 Differentiation and integration

In this section we deal with the series of derivatives and integrals given respectively by

$$\sum_{n=1}^{\infty} n a_n z^{n-1} \quad \text{and} \quad \sum_{n=0}^{\infty} a_n \frac{z^{n+1}}{n+1}$$

and show that in the interior of the disc of convergence the derivative and the integral of the sum are the sums of the series of derivatives and of integrals

$$D\left(\sum_{n=0}^{\infty} a_n z^n\right) = \sum_{n=1}^{\infty} n a_n z^{n-1}, \quad \int \sum_{n=0}^{\infty} a_n z^n dz = \sum_{n=0}^{\infty} a_n \frac{z^{n+1}}{n+1}.$$

In doing that we deal first with the real power series and then with the complex power series, since in the latter case, one needs to introduce suitable notions of differentiation and integration for functions of complex variables. We postpone the discussion of some partial results at the boundary to the next section.

a. Series of derivatives and of integrals

7.17 Proposition. *The power series $\sum_{n=0}^{\infty} a_n z^n$, $\sum_{n=1}^{\infty} n a_n z^{n-1}$, and $\sum_{n=0}^{\infty} a_n \frac{z^{n+1}}{n+1}$ all have the same radius of convergence.*

Proof. Let ρ and σ be respectively the radii of convergence of $\sum_{n=0}^{\infty} a_n z^n$ and of $\sum_{n=1}^{\infty} n a_n z^{n-1}$. Let us prove that $\rho = \sigma$. We first prove that $\sigma \leq \rho$. Assuming $\sigma > 0$ since otherwise the claim is trivial, fix z such that $|z| < \sigma$. Since

$$|a_n||z|^n \leq |z|n|a_n||z|^{n-1}, \quad n \geq 1,$$

we deduce that $\sum_{n=0}^{\infty} a_n z^n$ converges absolutely at z by the comparison test; thus $\rho \geq \sigma$.

Let us prove that $\rho \leq \sigma$. Assuming $\rho > 0$, for any fixed $|z| < \rho$, let r be such that $|z| < r < \rho$. We infer that $\sum_{n=0}^{\infty} |a_n| r^n$ converges, so that $\{|a_n| w^n\}$ is bounded, $|a_n| w^n \leq M$, hence

$$n|a_n||z|^{n-1} \leq M n \left(\frac{|z|}{|z| + \rho} \right)^{n-1}.$$

Therefore again the comparison test implies the absolute convergence of $\sum_{n=1}^{\infty} n a_n z^{n-1}$ at z ; z being arbitrary, we then conclude that $\sigma \geq \rho$.

Finally, $\sum_{n=0}^{\infty} a_n \frac{z^{n+1}}{n+1}$ has radius of convergence ρ , too, since $\sum_{n=0}^{\infty} a_n z^n$ is the series of derivatives of $\sum_{n=0}^{\infty} a_n \frac{z^{n+1}}{n+1}$. \square

7.18 ¶. Show that

$$\limsup_{n \rightarrow \infty} \sqrt[n]{(n+1)|a_{n+1}|} = \limsup_{n \rightarrow \infty} \sqrt[n]{\frac{|a_{n-1}|}{n}} = \limsup_{n \rightarrow \infty} \sqrt[n]{|a_n|}$$

inferring this way Proposition 7.17.

Iterating Proposition 7.17 we also conclude that, for every k , the series of the k -th derivatives

$$\sum_{n=k}^{\infty} n(n-1)(n-2) \cdots (n-k+1) a_n z^{n-k}$$

has the same radius of convergence of $\sum_{n=0}^{\infty} a_n z^n$.

b. Real power series

Consider a real power series $\sum_{n=0}^{\infty} a_n x^n$ in which $\{a_n\} \subset \mathbb{R}$ and $x \in \mathbb{R}$. Notice that its domain of convergence is an interval of radius ρ , ρ being the radius of convergence of the series, that can be either open, closed, left-closed or right-closed.

First we state the following simple theorem concerning the exchange of limit and integrals.

7.19 Theorem. Let $\{S_n\}$ be a sequence of functions $S_n : [a, b] \rightarrow \mathbb{R}$ that converges uniformly to $S : [a, b] \rightarrow \mathbb{R}$ on the bounded interval $[a, b]$. Then

$$\int_a^b S_n(x) dx \rightarrow \int_a^b S(x) dx.$$

Proof. This follows at once, observing that $S(x)$ being continuous by Theorem 7.13, we have

$$\begin{aligned} \left| \int_a^b (S_n(x) - S(x)) dx \right| &\leq (b-a) \sup_{x \in [a, b]} |S_n(x) - S(x)| \\ &= (b-a) \|S_n - S\|_{\infty, [a, b]}. \end{aligned} \quad (7.5)$$

□

7.20 Remark. Notice that both the assumptions of the uniform convergence and of performing the integrals on a bounded interval cannot be dropped in Theorem 7.19. For instance, starting with $\varphi(x) = xe^{-x}$,

- Choosing $S_n(x) := \frac{1}{n} \varphi(\frac{x}{n})$, we have $\|S_n\|_{\infty, [0, \infty[} = \frac{\|\varphi\|_{\infty, [0, \infty[}}{n} \rightarrow 0$, hence S_n converges uniformly to $S(x) = 0$, but

$$\int_0^{\infty} S_n(x) dx = \int_0^{\infty} \varphi(x) dx > 0 \quad \text{for all } n \geq 1$$

although for any $a > 0$,

$$\int_0^a S_n(x) dx = \int_0^{a/n} \varphi(x) dx \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

- Choosing $S_n(x) := n\varphi(nx)$, $x \in [0, 1]$, $S_n(x)$ converges pointwisely to 0 $\forall x \in [0, 1]$, and

$$\int_0^1 S_n(x) dx = \int_0^n \varphi(t) dt \rightarrow \int_0^\infty \varphi(t) dt > 0.$$

7.21 Theorem (Differentiation and integration of power series).

Let $S(x)$ be the sum of the power series $\sum_{n=0}^\infty a_n x^n$. If $\sum_{n=0}^\infty a_n x^n$ converges uniformly to S on an closed interval $[\alpha, \beta] \subset \mathbb{R}$, then

$$\int_\alpha^\beta S(x) dx = \sum_{n=0}^\infty a_n \int_\alpha^\beta x^n dx.$$

Consequently,

- (i) assuming that the radius of convergence ρ of $\sum_{n=0}^\infty a_n x^n$ is positive, we have

$$\int_0^x \sum_{n=0}^\infty a_n t^n dt = \sum_{n=0}^\infty a_n \frac{x^{n+1}}{n+1} \quad (7.6)$$

- for all x , $|x| < \rho$,
(ii) $S \in C^\infty(\cdot) - \rho, \rho[)$ and

$$D^k S(x) = \sum_{n=k}^\infty n(n-1) \cdots (n-k+1) x^{n-k}, \quad |x| < \rho. \quad (7.7)$$

Proof. The first claim and (i) follow at once from Theorem 7.19 since $S_n(x) := \sum_{j=0}^n a_j x^j$ converges uniformly to S .

- (ii) From (i) and Proposition 7.17 we get

$$\begin{aligned} \int_0^x \left(\sum_{n=1}^\infty n a_n t^{n-1} \right) dt &= \sum_{n=1}^\infty \left(n a_n \int_0^x t^{n-1} \right) dt \\ &= \sum_{n=1}^\infty a_n x^n = S(x) - a_0 = S(x) - S(0). \end{aligned}$$

By the fundamental theorem of calculus, S is then differentiable on $\{|x| < \rho\}$ and

$$S'(x) = \sum_{n=1}^\infty n a_n x^{n-1}, \quad |x| < \rho.$$

The proof is then easily completed by induction. □

7.22 ¶. Show the following

Proposition. Let $\{S_n\}$ be a sequence of functions $S_n : [a, b] \rightarrow \mathbb{R}$ of class $C^1[a, b]$. Suppose that the derivatives $S'_n : [a, b] \rightarrow \mathbb{R}$ converge uniformly to a function $T : [a, b] \rightarrow \mathbb{R}$ and that $S_n(x_0) \rightarrow \lambda \in \mathbb{R}$ as $n \rightarrow \infty$ for some $x_0 \in [a, b]$. Then

- (i) $\{S_n\}$ converges pointwisely to a function $S : [a, b] \rightarrow \mathbb{R}$,
(ii) S is differentiable on $[a, b]$, $S'(x) = T(x) \forall x \in [a, b]$ and S is of class $C^1([a, b])$.

[Hint: Compare (ii) of Theorem 7.21.]

7.23 Remark. Theorem 7.21 extends immediately to series $\sum_{n=0}^i i a_n x^n$ with $a_n \in \mathbb{C}$ and $x \in \mathbb{R}$. Assuming that $\sum_{n=0}^{\infty} a_n x^n$ has a positive radius of convergence $\rho > 0$, we have

$$\begin{aligned}\sum_{n=0}^{\infty} a_n x^n &= \sum_{n=0}^{\infty} \Re(a_n) x^n + i \sum_{n=0}^{\infty} \Im(a_n) x^n, \\ D\left(\sum_{n=0}^{\infty} a_n x^n\right) &= \left(\sum_{n=0}^{\infty} \Re(a_n) x^n\right) + i \left(\sum_{n=0}^{\infty} \Im(a_n) x^n\right), \\ \int_0^x \sum_{n=0}^{\infty} a_n t^n dt &= \int_0^x \sum_{n=0}^{\infty} \Re(a_n) t^n dt + i \int_0^x \sum_{n=0}^{\infty} \Im(a_n) t^n dt,\end{aligned}$$

for all $x \in \mathbb{R}$, $|x| < \rho$.

7.24 Example. Denote by $\text{Si}: [0, \infty[\rightarrow \mathbb{R}$ the function

$$\text{Si}(x) := \int_0^x \frac{\sin t}{t} dt.$$

By Theorem 7.21 we find

$$\int_0^x \frac{\sin t}{t} dt = \int_0^x \sum_{n=0}^{\infty} (-1)^n \frac{t^{2n}}{2n+1} dt = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)(2n+1)!}, \quad x > 0,$$

and the following error estimate in the approximation $\text{Si}(x) \sim \sum_{n=0}^p (-1)^n \frac{x^{2n+1}}{(2n+1)(2n+1)!}$ holds,

$$\left| \sum_{n=p}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)(2n+1)!} \right| \leq \frac{x^{2p+1}}{(2p+1)(2p+1)!}.$$

c. Power series and Taylor series

From (7.7) we infer the following.

7.25 Theorem. Let $\sum_{n=0}^{\infty} a_n x^n$ be a power series with positive radius of convergence. Then it is the Taylor series of its sum $S(x) := \sum_{n=0}^{\infty} a_n x^n$, that is,

$$a_n = \frac{D^n S(0)}{n!} \quad \forall n \geq 0. \quad (7.8)$$

Theorem 7.25 expresses the *rigidity of the sums of power series*: it suffices to know $S(x)$ in a small interval $] -\delta, \delta[$ of the real axis to know its derivatives at 0. By Theorem 7.25 all coefficients a_n are then identified, and in turn the sum $S(x)$ on the entire domain of convergence. Explicit formulas exist in some cases, but we do not pursue this point which is nowadays part of the *theory of functions of complex variables*.

Another immediate consequence is the following.

7.26 Theorem (Principle of identity of power series). Suppose that $\sum_{n=0}^{\infty} a_n x^n$ and $\sum_{n=0}^{\infty} b_n x^n$ converge in $] -\rho, \rho[$, $\rho > 0$, and that $\sum_{n=0}^{\infty} a_n x^n = \sum_{n=0}^{\infty} b_n x^n$ on $] -\rho, \rho[$. Then $a_n = b_n$ for all n .

d. Complex series

The theorem of differentiation and integration of series extends to complex power series provided suitable definitions of “complex derivative” and of “integral of a complex function” are given. We do not give here fully general definitions, as it would lead us into the *theory of functions of complex variables*. Here we make only a few remarks. Let $f : B \subset \mathbb{C} \rightarrow \mathbb{C}$ be a function defined on an open ball B centered at zero in the complex plane, $B := \{z \mid |z| < \rho\}$. We say that f has *complex derivative* at $z_0 \in B$ if the following limit exists in \mathbb{C} ,

$$\lim_{z \rightarrow z_0} \frac{f(z) - f(z_0)}{z - z_0} =: f'(z_0).$$

We define the *integral* of f from 0 to $z = x + iy \in B$ as

$$\int_0^z f(w) dw := \int_0^x f(t + i0) dt + \int_0^y f(x + it) dt.$$

7.27 Remark. Notice the following:

(i) We have

$$Dz^n := nz^{n-1}, \quad \int_0^z w^n dw := \frac{z^{n+1}}{n+1} \quad \forall z \in \mathbb{C}$$

as one can see by a direct computation.

(ii) Because of the fundamental theorem of calculus, if $f : B \rightarrow \mathbb{C}$ has a complex derivative that is continuous, then

$$f(z) - f(0) = \int_0^z Df(z) dz \quad (7.9)$$

as one can check from the definition.

(iii) From (ii) it follows: if f has a continuous complex derivative on B and $Df(z) = 0 \forall z \in B$, then f is constant on B .

(iv) If $g : B \rightarrow \mathbb{C}$ has a complex derivative, then the function $f : B \rightarrow \mathbb{C}$ defined by

$$f(z) := \int_0^z g(w) dw,$$

has a complex derivative on B and $Df(z) = g(z)$. This is actually a key point of the theory of functions of a complex variable.

With these definitions Theorem 7.21 extends to the following.

7.28 Theorem (Differentiation and integration of power series).

Let $\sum_{n=0}^{\infty} a_n z^n$ converge with a positive radius of convergence $\rho > 0$ and let $S(z)$ be its sum, $S(z) := \sum_{n=0}^{\infty} a_n z^n$. Then

(i) $S(z)$ has complex derivatives of any order in $\{|z| < \rho\}$ and

$$D^k S(z) = \sum_{n=k}^{\infty} n(n-1) \cdots (n-k+1) z^{n-k}, \quad |z| < \rho.$$

(ii)

$$\int_0^z \sum_{n=0}^{\infty} a_n w^n dw = \sum_{n=0}^{\infty} a_n \frac{z^{n+1}}{n+1}.$$

Proof. Assuming (iv) of Remark 7.27, one can repeat the proof of Theorem 7.21. Here we present a more direct proof.

(i) Fix z such that $|z| < \rho$ and let $\delta = \delta(z)$ be such that $|z| < \delta < \rho$. We notice that for any h with $0 < |h| < \delta - |z|$, we have

$$\begin{aligned} \frac{S(z+h) - S(z)}{h} &= \frac{1}{h} \left(\sum_{n=0}^{\infty} a_n (z+h)^n - \sum_{n=0}^{\infty} a_n z^n \right) \\ &= \frac{1}{h} \sum_{n=1}^{\infty} a_n \left((z+h)^n - z^n \right) \\ &= \frac{1}{h} \sum_{n=1}^{\infty} a_n \left(\sum_{k=0}^n \binom{n}{k} z^k h^{n-k} - z^n \right) \\ &= \sum_{n=1}^{\infty} n a_n \left(\sum_{k=0}^{n-1} z^k h^{n-k-1} \right) \\ &= \sum_{n=1}^{\infty} n a_n z^{n-1} + \sum_{n=2}^{\infty} a_n \left(\sum_{k=0}^{n-2} z^k h^{n-k-1} \right), \end{aligned}$$

that is,

$$\left| \frac{S(z+h) - S(z)}{h} - \sum_{n=1}^{\infty} n a_n z^{n-1} \right| \leq |\omega(z, h)|, \quad (7.10)$$

where $\omega(z, h) := \sum_{n=2}^{\infty} a_n \left(\sum_{k=0}^{n-2} \binom{n}{k} z^k h^{n-k-1} \right)$. We then estimate $\omega(z, h)$ by

$$\begin{aligned}
|\omega(z, h)| &\leq |h| \sum_{n=2}^{\infty} |a_n| \left(\sum_{k=0}^{n-2} \binom{n}{k} |z|^k |h|^{n-k-2} \right) \\
&= |h| \sum_{n=2}^{\infty} |a_n| \frac{n(n-1)}{2} \left(\sum_{k=0}^{n-2} \binom{n-2}{k} |z|^k |h|^{n-k-2} \right) \quad (7.11)
\end{aligned}$$

$$\begin{aligned}
&= |h| \sum_{n=2}^{\infty} |a_n| \frac{n(n-1)}{2} (|z| + |h|)^{n-2} \quad (7.12) \\
&\leq |h| \sum_{n=2}^{\infty} \frac{n(n-1)}{2} |a_n| \delta^n = C |h|
\end{aligned}$$

where $C \in \mathbb{R}$ is the sum of $\sum_{n=2}^{\infty} \frac{n(n-1)}{2} \delta^n$ which converges and does not depend on h . In conclusion, (7.10) and (7.11) yield $\frac{S(z+h)-S(z)}{h} \rightarrow \sum_{n=1}^{\infty} n a_n z^{n-1}$ as $h \rightarrow 0$. Since z has been chosen arbitrarily in the interior of the disc of convergence, we then conclude that S is differentiable at each point z with $|z| < \rho$, and

$$DS(z) = \sum_{n=1}^{\infty} n a_n z^{n-1}, \quad |z| < \rho.$$

By induction on k we then infer (i).

(ii) For $|z| < \rho$ and $t \in [0, 1]$, the complex series of the real variable $t \sum_{n=0}^{\infty} (a_n z^n) t^n$ has radius of convergence larger than 1 and sum $S(tz)$. From Remark 7.23 then

$$\int_0^z S(w) dw = z \int_0^1 S(tz) dt = z \sum_{n=0}^{\infty} a_n z^n \frac{1^{n+1}}{n+1} = \sum_{n=0}^{\infty} a_n \frac{z^{n+1}}{n+1}.$$

□

7.2 Further Results

7.2.1 Boundary values

Let $\sum_{n=0}^{\infty} a_n z^n$ be a power series with radius of convergence that we assume for the sake of simplicity to be 1. As we have seen, if the series converges absolutely at some boundary point z , $|z| = 1$, then it converges absolutely at every boundary point.

In some cases the following test is useful.

7.29 Proposition (Absolute convergence at the boundary). Let $\sum_{n=0}^{\infty} a_n z^n$ be a power series with a positive radius of convergence $\rho > 0$ and sum $S(z)$. Suppose that for some $z_0 \in \mathbb{C}$ with $|z_0| = \rho$ and for each integer n , $a_n z_0^n$ is a nonnegative real number and that $\limsup_{r \rightarrow 1^-} S(r z_0) < +\infty$. Then $\sum_{n=0}^{\infty} a_n z^n$ converges absolutely at all z such that $|z| = \rho$.

Proof. In fact, if $|z| = |z_0| = \rho$, for any $p \geq 0$ we have $\sum_{n=0}^p a_n z_0^n r^n \leq \sum_{n=0}^{\infty} a_n z_0^n r^n$, hence

$$\begin{aligned} \sum_{n=0}^p |a_n| |z|^n &= \sum_{n=0}^p |a_n| |z_0|^n = \lim_{r \rightarrow 1^-} \sum_{n=0}^p a_n z_0^n r^n \\ &\leq \limsup_{r \rightarrow 1^-} \sum_{n=0}^{\infty} a_n (r z_0)^n = \limsup_{r \rightarrow 1^-} S(r z_0) < +\infty. \end{aligned}$$

□

Let $\sum_{n=0}^{\infty} a_n z^n$ be a power series with radius of convergence 1. If the series $\sum_{n=0}^{\infty} a_n z^n$ does not converge at every boundary point $|z| = 1$, then at those points at which it converges it does not converge absolutely. The following theorem deals with one such case.

7.30 Theorem (Dirichlet). Let $\sum_{j=0}^n a_n z^n$ be a series with radius of convergence 1. Suppose that the sequence $\{a_n\}$ converges to zero and has bounded total variation,

$$\sum_{n=0}^{\infty} |a_{n+1} - a_n| < \infty.$$

Then $\sum_{n=0}^{\infty} a_n z^n$ converges for all z with $|z| = 1$ and $z \neq 1$ and

$$\left| \sum_{n=p}^{\infty} a_n z^n \right| \leq \frac{4}{|1-z|} \sum_{n=p}^{\infty} |a_{p+1} - a_p|, \quad (7.13)$$

in particular $\sum_{n=0}^{\infty} a_n z^n$ converges uniformly on every domain $D_\rho := \{z \mid |z| \leq 1, |1-z| \geq \rho\}$ for all $\rho > 0$. If moreover $S(z)$ is its sum, then $S(z)$ is continuous on $\{z \mid |z| \leq 1, z \neq 1\}$ and

$$(1-z)S(z) \rightarrow 0 \quad \text{as } z \rightarrow 1, |z| \leq 1.$$

Proof. For $|z| < 1$ we have

$$\left| \sum_{j=0}^n z^j \right| = \left| \frac{1-z^{j+1}}{1-z} \right| \leq \frac{2}{|1-z|};$$

therefore by Dirichlet's test, Theorem 6.55, $\sum_{n=0}^{\infty} a_n z^n$ converges and (7.13) holds. We therefore infer from (7.13) the uniform convergence of

$\sum_{n=0}^{\infty} a_n z^n$ on D_ρ . This proves the first part of the theorem, and the continuity of the sum $S(z)$ on $\{z \mid |z| \leq 1, z \neq 1\}$.

To prove the second part, for $\epsilon > 0$ choose n in such a way that $\sum_{j=n+1}^{\infty} |a_{j+1} - a_j| < \epsilon$. If $S_n(z)$ denotes the n -th partial sum, by (7.13) we have for $|z| \leq 1$,

$$\begin{aligned} |(1-z)S(z)| &\leq |(1-z)S_n(z)| + |(1-z)(S(z) - S_n(z))| \\ &= |(1-z)S_n(z)| + \left| (1-z) \sum_{j=n+1}^{\infty} a_n z^n \right| \\ &\leq |(1-z)S_n(z)| + 4\epsilon. \end{aligned}$$

Being that $(1-z)S_n(z)$ is a polynomial, hence a continuous function, there exists $\delta > 0$ such that $|(1-z)S_n(z)| < \epsilon$ for $|z-1| < \delta$, hence we conclude for $|z-1| < \delta$ and $|z| \leq 1$ that $|(1-z)S(z)| \leq 5\epsilon$. \square

Suppose that $\sum_{n=0}^{\infty} a_n z^n$ converges at a point z_0 of the boundary $\{z \mid |z| = 1\}$ of its disc of convergence; is its sum continuous at z_0 ? Of course the answer is yes if $\sum_{n=0}^{\infty} a_n z^n$ converges absolutely; in this case, in fact, it converges uniformly in $\{z \mid |z| \leq 1\}$ by Theorems 7.14 and 7.13. The next theorem gives a partial answer in the general case.

7.31 Theorem (Abel). *Suppose $\sum_{n=0}^{\infty} a_n z^n$ has radius of convergence 1 and converges at z_0 with $|z_0| = 1$. Then the series of real powers with complex coefficients $\sum_{n=0}^{\infty} (a_n z_0^n) t^n$ converges uniformly on $[0, 1]$, and its sum $s(t) := S(tz_0)$, $S(z)$ being the sum of $\sum_{n=0}^{\infty} a_n z^n$, is continuous on $[0, 1]$.*

Proof. Set $\alpha_n := t^n$, $\beta_n := a_n z_0^n$ and $B_n := \sum_{j=0}^n \beta_j$. By assumption $\sum_{n=0}^{\infty} \beta_n$ converges and

$$\sum_{n=0}^{\infty} |t^{n+1} - t^n| = \begin{cases} 1 & \text{if } 0 < t < 1, \\ 0 & \text{if } t = 1, \end{cases}$$

since $t \in [0, 1]$. By Abel's test, Theorem 6.57, applied with $a_n := \alpha_n$ and $b_n := \beta_n$, we get

$$\begin{aligned} \left| \sum_{j=n+1}^{\infty} \alpha_j \beta_j \right| &\leq \sup_{j \geq n+1} |B_j - B_{p-1}| \left\{ 2 \sum_{j=n+1}^{\infty} |t^j - t^{j+1}| + |t^p| \right\} \\ &\leq 3 \sup_{j \geq n+1} |B_j - B_{p-1}|, \end{aligned} \quad (7.14)$$

where $B_p := \sum_{n=0}^p \beta_p$. On the other hand, given $\epsilon > 0$, there exists \bar{n} such that $|B_p - B_n| < \epsilon$ for $n, p \geq \bar{n}$, hence

$$\left| \sum_{j=n+1}^{\infty} a_n z_0^n t^n \right| = \left| \sum_{j=n+1}^{\infty} \alpha_j \beta_j \right| \leq 3\epsilon.$$

Therefore the series $\sum_{n=0}^{\infty} (a_n z_0^n) t^n$ converges uniformly in $[0, 1]$, and the function $s(t) := \sum_{n=0}^{\infty} (a_n z_0^n) t^n$ is continuous in $[0, 1]$ by Theorem 7.13. \square

In particular, we can state the following.

7.32 Corollary. *Let $\sum_{n=0}^{\infty} a_n z^n$ be a power series with radius of convergence $\rho > 0$. If it converges at z_0 with $|z_0| = \rho$, then its sum is continuous when restricted on the closed segment joining 0 with z_0 . Moreover we have*

$$\int_0^{z_0} S(w) dw = z_0 \int_0^1 S(tz_0) dt = \sum_{n=0}^{\infty} a_n \frac{z_0^{n+1}}{n+1}. \quad (7.15)$$

As a consequence we can prove the claim (i) in Remark 6.61.

7.33 Theorem (Abel). *If $\sum_{n=0}^{\infty} a_n$ and $\sum_{n=0}^{\infty} b_n$ converge respectively to A and B and the product series $\sum_{n=0}^{\infty} c_n$, $c_n := \sum_{k=0}^n a_k b_{n-k}$, converges to C , then $C = AB$.*

Proof. Set $f(x) = \sum_{n=0}^{\infty} a_n x^n$, $g(x) = \sum_{n=0}^{\infty} b_n x^n$ and $h(x) = \sum_{n=0}^{\infty} c_n x^n$, $0 \leq x \leq 1$. Since $f(x)g(x) = h(x)$ for $0 \leq x < 1$, see Theorem 6.60, and $f(x) \rightarrow A$, $g(x) \rightarrow B$ and $h(x) \rightarrow C$ by Theorem 7.31, the claim follows. \square

7.2.2 Product and composition of power series

An immediate consequence of Theorem 6.60 is the following.

7.34 Theorem. *Let $\sum_{n=0}^{\infty} a_n z^n$ and $\sum_{n=0}^{\infty} b_n z^n$ be two power series with respectively radii of convergence $\rho_a > 0$ and $\rho_b > 0$. Then the series $\sum_{n=0}^{\infty} c_n z^n$, where $c_n := \sum_{k=0}^n a_k b_{n-k}$, has radius of convergence $\rho_c \geq \max(\rho_a, \rho_b)$ and*

$$\sum_{n=0}^{\infty} c_n z^n = \left(\sum_{n=0}^{\infty} a_n z^n \right) \left(\sum_{n=0}^{\infty} b_n z^n \right)$$

for all z , $|z| \leq \min(\rho_a, \rho_b)$.

Notice that ρ_c is at least the *maximum* of ρ_a and ρ_b . For instance, if $\sum_{n=0}^{\infty} b_n z^n$ is a polynomial, then $\sum_{n=0}^{\infty} c_n z^n$ is a polynomial, too, hence $\rho_b = \rho_c = +\infty$, no matter the size of ρ_a .

a. Weierstrass's double series theorem

Suppose that all series

$$S_m(z) := \sum_{l=0}^{\infty} a_{ml} z^l, \quad m = 0, 1, 2, \dots,$$

are convergent at least for $|z| < R$, $R > 0$ and that for every $\rho < R$

$$S(z) := \sum_{m=0}^{\infty} S_m(z)$$

is uniformly convergent on $|z| \leq \rho$. Then we have

7.35 Theorem. *The coefficients of the power series of z in the respective series form a convergent series and if we set*

$$a_k := \sum_{m=0}^{\infty} a_{mk},$$

the series $\sum_{k=0}^{\infty} a_k z^k$ sums to $S(z)$ at least for $|z| < R$. Moreover $S(z)$ is infinitely differentiable and

$$\frac{1}{n!} D^{(n)} S(0) = \sum_{m=0}^{\infty} D^{(n)} S_m(0).$$

This follows at once from Proposition 6.66

7.2.3 Taylor series: examples

Taylor series of given smooth functions (see Section 6.1) are important examples of power series on which one can test the theory.

7.36 Example. We proved in Example 6.15 that

$$\sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x, \quad \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} = \sin x, \quad \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n}}{(2n)!} = \cos x$$

for all $x \in \mathbb{R}$, thus the radius of convergence of all these series is ∞ . Convergence to their respective sum is uniform on every *bounded* interval.

7.37 Example (Geometric series). We saw at several places that the geometric series $\sum_{n=1}^{\infty} x^n$ converges if and only if $|x| < 1$, and sums to

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}, \quad |x| < 1.$$

The convergence is uniform on every interval $[-r, r]$ with $r < \rho$. Observe that the convergence is not uniform, neither in the open interval $] - 1, 1[$ nor in the interval $] - 1, 0]$ in which the sum is bounded, since the quantity

$$\left\| \frac{1}{1-x} - \sum_{k=0}^n x^k \right\|_{[\infty,]-1, 0]} = \sup_{x \in]-1, 0]} \left| \frac{x^{n+1}}{1-x} \right| = 1.$$

7.38 Example (Logarithm). Replacing x by $-x$ in

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n, \quad |x| < 1,$$

we get

$$\frac{1}{1+x} = \sum_{n=0}^{\infty} (-1)^n x^n, \quad |x| < 1,$$

and integrating,

$$\log(1+x) = \int_0^x \frac{1}{1+t} dt = \int_0^x \sum_{n=0}^{\infty} (-1)^n t^n dt = \sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1}, \quad |x| < 1, \quad (7.16)$$

an equality that we already know by an ad hoc computation (see Example 6.13). In fact there we proved that $\sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1}$ converges if $-1 < x \leq 1$, does not converge if $x = -1$, and the equality (7.16) holds if $-1 < x \leq 1$.

On the other hand, the radius of convergence of $\sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1}$ is 1 since $1/\rho = \limsup_{n \rightarrow \infty} \sqrt[n]{1/n} = 1$. We then conclude that $\sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1}$ converges if and only if $-1 < x \leq 1$ and

$$\sum_{n=0}^{\infty} (-1)^n \frac{x^{n+1}}{n+1} = \log(1+x), \quad x \in]-1, 1].$$

7.39 Example (Arc tangent). Replacing x by $-x^2$ in

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n, \quad |x| < 1,$$

and integrating according to Theorem 7.19 we get

$$\arctan x = \int_0^x \frac{1}{1+t^2} dt = \int_0^x \sum_{n=0}^{\infty} (-1)^n t^{2n} dt = \sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1}, \quad |x| < 1, \quad (7.17)$$

see Example 6.14. There we proved the convergence of the series and the equality (7.17) also for $x = \pm 1$.

We can prove the result at $x = \pm 1$ by means of the theory. In fact, if $x = \pm 1$ the series reduces to $\pm \sum_{n=1}^{\infty} (-1)^n \frac{1}{2n+1}$ which converges by the Leibniz test. Then Abel's theorem yields continuity of the sum of the series at ± 1 , thus, passing to the limit in both sides of (7.17) as $x \rightarrow \pm 1$, we infer

$$\arctan \pm 1 = \pm \sum_{n=0}^{\infty} (-1)^n \frac{(\pm 1)^{2n+1}}{2n+1}.$$

On the other hand, since $|x|^{2n+1}/(2n+1) \rightarrow +\infty$ if $|x| > 1$, we conclude that $\sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1}$ converges if and only if $|x| \leq 1$, and

$$\sum_{n=0}^{\infty} (-1)^n \frac{x^{2n+1}}{2n+1} = \arctan x, \quad |x| \leq 1.$$

7.40 Example (The binomial series). We claim that

$$(1+x)^\alpha = \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n, \quad \alpha \in \mathbb{R}, \quad |x| < 1, \quad (7.18)$$

where

$$\binom{\alpha}{n} := \begin{cases} 1 & \text{if } n = 0, \\ \frac{\alpha(\alpha-1)(\alpha-2)\cdots(\alpha-n+1)}{n!} & \text{if } n \geq 1. \end{cases}$$

Notice that $D^n(1+x)^\alpha = \alpha(\alpha-1)\cdots(\alpha-n+1)(1+x)^{\alpha-n}$, hence the series in (7.18) is the Taylor series of $(1+x)^\alpha$ centered at zero.

Since

$$\frac{\left| \binom{\alpha}{n+1} \right|}{\left| \binom{\alpha}{n} \right|} = \frac{|\alpha - n|}{|\alpha - n + 1|} \rightarrow 1 \quad \text{as } n \rightarrow \infty$$

we infer, see Example 2.57, $\sqrt[n]{|a_n|} \rightarrow 1$; therefore the series in (7.18) has radius of convergence 1.

Let $S(x) := \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n$, $|x| < 1$. By differentiating we then find, similarly to 5.53 of [GM1],

$$(1+x)S'(x) = \alpha S(x), \quad |x| < 1,$$

hence

$$\left(\frac{S(x)}{(1+x)^\alpha} \right)' = \frac{(1+x)S'(x) - \alpha S(x)}{(1+x)^{\alpha+1}} = 0.$$

Therefore we conclude that $S(x) = c(1+x)^\alpha$ for $|x| < 1$, c being a constant; finally from $S(0) = 1$ we infer $c = 1$.

7.41 Example (The arc sine). Replacing x with $-x^2$ and choosing $\alpha = -1/2$, in (7.18) we get

$$\frac{1}{\sqrt{1-x^2}} = \sum_{n=0}^{\infty} \binom{-1/2}{n} (-1)^n x^{2n}, \quad |x| < 1$$

and, integrating, we get

$$\begin{aligned} \arcsin x &= \int_0^x \frac{1}{\sqrt{1-t^2}} dt = \int_0^x \sum_{n=0}^{\infty} (-1)^n \binom{-1/2}{n} t^{2n} dt \\ &= \sum_{n=0}^{\infty} (-1)^n \binom{-1/2}{n} \frac{x^{2n+1}}{2n+1} = \sum_{n=0}^{\infty} \frac{(2n-1)!!}{(2n)!!} \frac{x^{2n+1}}{2n+1} \end{aligned} \quad (7.19)$$

for $|x| < 1$, and that the series in (7.19) has radius of convergence 1. The series in (7.19) actually converges absolutely if $|z| = 1$, hence uniformly in $\{z \mid |z| \leq 1\}$. In fact, the coefficients of the series in (7.19), that we denote by $\{c_n\}$, are nonnegative, hence for all $p \geq 1$,

$$\sum_{n=0}^p |c_n| = \sum_{n=0}^p c_n = \lim_{r \rightarrow 1^-} \sum_{n=1}^p c_n r^n \leq \lim_{r \rightarrow 1^-} \sum_{n=0}^{\infty} c_n r^n = \lim_{r \rightarrow 1^-} \arcsin r = \frac{\pi}{2}.$$

7.42 Example. Similarly, choosing in (7.18) $\alpha = 1/2$, we get

$$\sum_{n=0}^{\infty} \binom{1/2}{n} (-1)^n x^n = \sqrt{1-x}, \quad |x| < 1$$

and the series has a radius of convergence 1. Actually, $\sum_{n=0}^{\infty} c_n z^n$, $c_n := \binom{-1/2}{n} (-1)^n$, converges absolutely if $|z| = 1$, hence uniformly in $\{z \mid |z| \leq 1\}$. We in fact have $c_n < 0$ $\forall n \geq 1$, hence, for $p \geq 1$,

$$\begin{aligned}\sum_{n=0}^p |c_n| &= 1 - \sum_{n=1}^p c_n = 2 - \lim_{r \rightarrow 1^-} \sum_{n=0}^p c_n r^n \\ &\leq 2 - \lim_{r \rightarrow 1^-} \sum_{n=0}^{\infty} \binom{-1/2}{n} (-r)^n = \lim_{r \rightarrow 1^-} 2 - \sqrt{1-r} = 2.\end{aligned}$$

7.43 Example. The series $\sum_{n=0}^{\infty} n^2 z^n$ has radius of convergence 1. Writing $n^2 z^n = n(n-1)z^n + nz^n = z^2 n(n-1)z^{n-2} + znz^{n-1}$, and summing, we get, for $|z| < 1$,

$$\sum_{n=0}^{\infty} n^2 z^n = z^2 D^2 \left(\sum_{n=0}^{\infty} z^n \right) + zD \left(\sum_{n=0}^{\infty} z^n \right) = z^2 D^2 \frac{1}{1-z} + zD \frac{1}{1-z} = \dots$$

7.44 Example. We compute

$$\sum_{n=2}^{\infty} \frac{n+2}{n-1} z^n.$$

Writing

$$\frac{n+2}{n-1} = 1 + \frac{3}{n-1},$$

multiplying by z^n and summing we get

$$\sum_{n=2}^{\infty} \frac{n+2}{n-1} z^n = \sum_{n=2}^{\infty} z^n + 3z \sum_{n=2}^{\infty} \frac{z^{n-1}}{n-1} = \frac{z^2}{1-z} - 3z \log(1-z),$$

by using the identities

$$\sum_{n=0}^{\infty} z^n = \frac{1}{1-z}, \quad \sum_{n=0}^{\infty} \frac{z^{n+1}}{n+1} = -\log(1-z), \quad |z| < 1.$$

7.45 Example. We compute

$$\sum_{n=3}^{\infty} \frac{1}{(n+1)(n-2)} z^n.$$

Since

$$\frac{1}{(n+1)(n-2)} = \frac{1}{3} \frac{1}{n+1} - \frac{1}{3} \frac{1}{n-2},$$

multiplying by z^n and summing, we get, for $|z| < 1$,

$$\begin{aligned}\sum_{n=3}^{\infty} \frac{1}{(n+1)(n-2)} z^n &= \frac{1}{3z} \sum_{n=3}^{\infty} \frac{z^{n+1}}{n+1} - \frac{z^2}{3} \sum_{n=3}^{\infty} \frac{z^{n-2}}{n-2} \\ &= \frac{1}{3z} \left(-\log(1-z) - z - \frac{z^2}{2} - \frac{z^3}{3} \right) + \frac{z^2}{3} \log(1-z).\end{aligned}$$

Here we used that $\sum_{n=0}^{\infty} \frac{z^{n+1}}{n+1} = -\log(1-z)$ if $|z| < 1$.

7.3 Some Applications

In this section we illustrate some applications of the theory of power series.

7.46 Example. Conversely we may express $\sum_{n=1}^{\infty} x^n/n^2$ as an integral. In fact for $|x| < 1$ we have

$$\sum_{n=1}^{\infty} \frac{x^n}{n^2} = \sum_{n=1}^{\infty} \frac{1}{n} \int_0^x t^{n-1} dt = \int_0^x \sum_{n=1}^{\infty} \frac{t^{n-1}}{n} dt = \int_0^x \frac{1}{t} \sum_{n=1}^{\infty} \frac{t^n}{n} dt = \int_0^x \frac{\log(1-t)}{t} dt.$$

This is easily justified since all series in the previous formulas have radius of convergence 1 and therefore converge uniformly on $[0, x]$ (respectively $[x, 0]$ if $x < 0$) if $|x| < 1$.

7.3.1 Complex functions

7.47 Complex exponential. We defined in (4.6) the complex exponential e^z by

$$e^z := e^x(\cos y + i \sin y), \quad z =: x + iy \in \mathbb{C}.$$

Proposition. We have

$$e^z := \sum_{n=0}^{\infty} \frac{z^n}{n!} \quad \forall z \in \mathbb{C}. \quad (7.20)$$

Proof. Since the series of $\sin y$ and $\cos y$ converge in \mathbb{R} , we get

$$\begin{aligned} e^{iy} = \cos y + i \sin y &= \sum_{n=0}^{\infty} (-1)^n \frac{y^{2n}}{(2n)!} + i \sum_{n=0}^{\infty} (-1)^n \frac{y^{2n+1}}{(2n+1)!} \\ &= \sum_{n=0}^{\infty} \frac{(iy)^{2n}}{(2n)!} + \sum_{n=0}^{\infty} \frac{(iy)^{2n+1}}{(2n+1)!} = \sum_{n=0}^{\infty} \frac{(iy)^n}{n!}, \end{aligned}$$

i.e., our claim for $z = iy$, $y \in \mathbb{R}$. Therefore by Theorem 6.60 we infer

$$e^{x+iy} = e^x e^{iy} = \sum_{n=0}^{\infty} \frac{x^n}{n!} \sum_{n=0}^{\infty} \frac{(iy)^n}{n!} = \sum_{n=0}^{\infty} c_n$$

where

$$c_n := \sum_{k=0}^n \frac{x^k}{k!} \frac{(iy)^{n-k}}{(n-k)!} = \frac{1}{n!} \sum_{k=0}^n \binom{n}{k} x^k (iy)^{n-k} = \frac{1}{n!} (x + iy)^n = \frac{z^n}{n!}.$$

□

The complex differentiation theorem for series yields also

$$De^z = \sum_{n=0}^{\infty} n \frac{z^{n-1}}{n!} = \sum_{n=1}^{\infty} \frac{z^{n-1}}{(n-1)!} = e^z.$$

7.48 The complex logarithm. We recall that the principal determination of the logarithm can be written as

$$\log z := \log |z| + i \arg z, \quad z \neq 0,$$

where $\arg z \in [-\pi, \pi[$ and $\arg 1 = 0$.

Proposition. We have

$$\log(1+z) = \sum_{n=0}^{\infty} (-1)^n \frac{z^{n+1}}{n+1}, \quad |z| < 1. \quad (7.21)$$

Proof. We observe that for $z \neq 0$,

$$e^{\log z} = e^{\log |z|} e^{i \arg z} = z$$

and the function $\log z$ is continuous on $\{z = x + iy \mid y = 0, x \leq 0\}$. Similar to the proof of the differentiability of the inverse of a real function (see Theorem 4.16 of [GM1]), one sees that $\log z$ is differentiable at the points at which it is continuous, i.e., on $\{z = x + iy \mid y \neq 0, x \leq 0\}$, and

$$D \log z = \frac{1}{z}, \quad z \in \{z = x + iy \mid y \neq 0, x \leq 0\}.$$

Integrating, see Remark 7.27,

$$\begin{aligned} \log(1+z) - \log 1 &= \int_0^z \frac{dz}{z+1} = \int_0^z \sum_{n=0}^{\infty} (-1)^n z^n dz \\ &= \sum_{n=0}^{\infty} (-1)^n \frac{z^{n+1}}{n+1} dz, \quad |z| < 1. \end{aligned}$$

□

7.49 ¶. Show that equality (7.21) holds for all z with $|z| = 1$, $z \neq -1$. [*Hint.* Use Abel's theorem and the continuity of $\log(1+z)$.]

7.50 Complex trigonometric and hyperbolic functions. Starting from the complex exponential one defines the complex sine and cosine, and the complex hyperbolic sine and cosine by

$$\begin{aligned} \sin z &= \frac{e^{iz} - e^{-iz}}{2i}, & \cos z &= \frac{e^{iz} + e^{-iz}}{2}, \\ \sinh z &= \frac{e^z - e^{-z}}{2}, & \cosh z &= \frac{e^z + e^{-z}}{2} \end{aligned}$$

that actually means by (7.20)

$$\cos z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!}, \quad \sin z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)!},$$

$$\cosh z = \sum_{n=0}^{\infty} \frac{z^{2n}}{(2n)!}, \quad \sin z = \sum_{n=0}^{\infty} \frac{z^{2n+1}}{(2n+1)!}.$$

Trivially the restrictions of the four previous functions to the real axis agree with the corresponding real functions. It is possible to derive several complex “trigonometric” identities, that are formally equivalent to the real ones:

$$\begin{aligned} \sin^2 z + \cos^2 z &= 1, & e^{iz} &= \cos z + i \sin z, \\ \sin(-z) &= -\sin z, & \cos(-z) &= \cos z, \\ \cosh^2 z - \sinh^2 z &= 1, & e^z &= \cosh z + \sinh z, \\ \sinh(-z) &= -\sinh z, & \cosh(-z) &= \cosh z, \\ \cosh(iz) &= \cos z, & \sinh(iz) &= i \sin z. \end{aligned}$$

and

$$\begin{aligned} \cos(z+w) &= \cos z \cos w - \sin z \sin w, \\ \sin(z+w) &= \cos z \sin w + \cos w \sin z, \\ \cosh(z+w) &= \cosh z \cosh w + \sinh z \sinh w, \\ \sinh(z+w) &= \cosh z \sinh w + \cosh w \sinh z. \end{aligned}$$

Consequently

$$\begin{aligned} \sin z &= 2 \sin(z/2) \cos(z/2), \\ \cos^2(z/2) &= (1 + \cos z)/2, \\ \sin w - \sin z &= 2 \cos\left(\frac{w+z}{2}\right) \sin\left(\frac{w-z}{2}\right), \\ \cos w - \cos z &= -2 \sin\left(\frac{w+z}{2}\right) \sin\left(\frac{w-z}{2}\right), \\ &\dots \end{aligned}$$

7.3.2 An alternate definition of π , e and of elementary functions

In [GM1] we defined e , π , the exponential function e^x and the trigonometric functions $\sin x$ and $\cos x$ using several tricks that involve the infinitesimal calculus. Here we want to point out that all these facts can be subsumed by the power series

$$\sum_{n=0}^{\infty} \frac{z^n}{n!}$$

that converges absolutely in \mathbb{C} . Define the *complex exponential function* as

$$\exp z := \sum_{n=0}^{\infty} \frac{z^n}{n!} \quad \forall z \in \mathbb{C}.$$

With some work, using several theorems, including Cauchy’s theorem about product of series, and the complex differentiation of the sums of complex power series, one may prove

(i) ADDITION THEOREM. $\exp(z+w) = (\exp z)(\exp w)$,

- (ii) $|\exp z| = \exp(\Re z)$, $|\exp z| = 1$ iff $z = iy$, $y \in \mathbb{R}$,
 (iii) $\exp z$ is differentiable in the complex sense infinitely many times with derivatives of any order equal to $\exp z$.

Moreover, if $z = x + i0$ is real, then we have for $\exp x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$,

$$\begin{cases} D(\exp(x)) = \exp(x), \\ \exp(0) = 1. \end{cases}$$

In other words the restriction of the exponential to the real axis is the exponential function we already know. At this point we introduce the two complex functions

$$\cos z := \frac{\exp(iz) + \exp(-iz)}{2}, \quad \sin z := \frac{\exp(iz) - \exp(-iz)}{2i}. \quad (7.22)$$

From (7.22) we obtain $\exp(iz) = \cos z + i \sin z$, $z \in \mathbb{C}$, hence, formally, the Euler identity,

$$\exp(it) = \cos t + i \sin t.$$

Again from (7.22), we infer $D \sin z = \cos z$, $D \cos z = -\sin z$, from which, the functions of real variable $y(x) := \sin(x + i0)$ and $z(x) := \cos(x + i0)$ are respectively solutions of Cauchy problems

$$\begin{cases} y'' + y = 0, \\ y(0) = 0, \quad y'(0) = 1, \end{cases} \quad \text{and} \quad \begin{cases} y'' + y = 0, \\ y(0) = 1, \quad y'(0) = 0. \end{cases}$$

In other words, $x \rightarrow \sin(x + i0)$ and $x \rightarrow \cos(x + i0)$ are the trigonometric functions that we already know. Again from (7.22), we get

$$\cos z := \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!}, \quad \sin z := \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)!}, \quad (7.23)$$

which furnishes the needed developments.

Recovering the number π , and discussing the periodicity of $\sin x$ and $\cos x$, $x \in \mathbb{R}$, from the complex exponential is a little tricky, but it can be done. One starts proving that $\exp : \mathbb{C} \rightarrow \mathbb{C} \setminus \{0\}$ is onto, and then one observes that $\exp : \mathbb{C} \rightarrow \mathbb{C} \setminus \{0\}$ is a homomorphism from the additive group \mathbb{C} into the multiplicative group $\mathbb{C} \setminus \{0\}$ which is onto but not injective: its kernel is given by

$$\ker(\exp) = \{w \in \mathbb{C} \mid \exp w = 1\}.$$

At this point one can show⁴ that there exists a unique positive number, that we call π , such that

$$\ker(\exp) = 2\pi i\mathbb{Z}, \quad \text{i.e.,} \quad \exp(2\pi k) = 1 \quad \forall k \in \mathbb{Z}.$$

The addition formulas for the sine and the cosine yield the 2π -periodicity of $\sin x$ and $\cos x$.

Concluding, we may regard π as one of

$$\text{zeros of } \sin z = k\pi, \quad k \in \mathbb{Z},$$

$$\text{zeros of } \cos z = \frac{\pi}{2} + k\pi, \quad k \in \mathbb{Z},$$

$$\text{periods of } \exp z = \ker(\exp z) = 2k\pi, \quad k \in \mathbb{Z},$$

$$\text{periods of } \sin z = 2\pi\mathbb{Z},$$

$$\text{periods of } \cos z = 2\pi\mathbb{Z}.$$

⁴ See, e.g., the paper by E. Remmert in H.E. Ebbinghens, H. Hermes, F. Hirzebruch, M. Koecher, K. Meier, J. Neurich, A. Prestel, R. Remmert, *Numbers*, Springer, New York, 1988.

7.3.3 Series solutions of differential equations

Power series turn out to be very useful in solving ODEs. Without entering the question of when or if an ODE has a series solution or whether all solutions can be represented as a power series, we confine ourselves here to presenting a few examples.

7.51 Example. Suppose that the equation

$$y'' - y = 0$$

has a solution of the form

$$y(x) = \sum_{k=0}^{\infty} a_k x^k$$

with a positive radius of convergence. We therefore have

$$0 = \sum_{k=2}^{\infty} k(k-1)a_k x^{k-2} - \sum_{k=0}^{\infty} a_k x^k = \sum_{k=2}^{\infty} (k(k-1)a_k - a_{k-2})x^{k-2},$$

hence, by the principle of identity of series,

$$k(k-1)a_k = a_{k-2}, \quad k = 2, 3, 4, \dots$$

For k even, $k = 2n$, $n \geq 1$, we then find

$$a_{2n} = \frac{a_{2n-2}}{2n(2n-1)}, \quad \text{i.e.,} \quad a_{2n} = \frac{a_0}{(2n)!},$$

and, for k odd, $k = 2n+1$, $n \geq 1$,

$$a_{2n+1} = \frac{a_{2n-1}}{(2n+1)2n}, \quad \text{i.e.,} \quad a_{2n+1} = \frac{a_1}{(2n+1)!}.$$

Since the series $\sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!}$ and $\sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!}$ converge on \mathbb{R} , we conclude that

$$y(x) = a_0 \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!} + a_1 \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!} = a_0 \cosh x + a_1 \sinh x, \quad x \in \mathbb{R}$$

is a solution of $y'' - y = 0$.

7.52 Example. Similarly, for the equation

$$y'' - xy = 0,$$

assuming that the series $\sum_{k=0}^{\infty} a_k x^k$ with positive radius of convergence is a solution of the ODE, we find

$$2a_2 + \sum_{k=1}^{\infty} \left((k+2)(k+1)a_{k+2} - a_{k-1} \right) x^k = 0,$$

i.e.,

$$a_2 = 0, \quad (k+2)(k+1)a_{k+2} - a_{k-1} = 0, \quad k = 1, 2, 3, \dots,$$

which yield

$$\begin{cases} a_{3n} = \frac{a_0}{2 \cdot 3 \cdot 5 \cdot 6 \cdots (3n-1)3n}, & n = 1, 2, \dots, \\ a_{3n+1} = \frac{a_1}{3 \cdot 4 \cdot 6 \cdot 7 \cdots 3n(3n+1)}, & n = 1, 2, \dots, \\ a_{3n+2} = 0, & n = 0, 1, 2, \dots \end{cases}$$

We find again two series which converge for all $x \in \mathbb{R}$,

$$y_1(x) := 1 + \sum_{n=1}^{\infty} \frac{x^{3n}}{2 \cdot 3 \cdot 5 \cdot 6 \cdots (3n-1)3n},$$

$$y_2(x) := 1 + \sum_{n=1}^{\infty} \frac{x^{3n+1}}{3 \cdot 4 \cdot 6 \cdot 7 \cdots 3n(3n+1)}$$

and such that

$$y(x) = a_0 y_1(x) + a_1 y_2(x)$$

solves the equation for all a_0 and a_1 .

7.53 Example. Consider the equation

$$x^2 y'' + (x^2 + x)y' - y = 0$$

and suppose that the series $\sum_{k=0}^{\infty} a_k x^k$ has a positive radius of convergence and is a solution of the equation. In this case

$$\begin{aligned} 0 &= \sum_{k=0}^{\infty} k(k-1)a_k x^k + \sum_{k=0}^{\infty} k a_k x^{k+1} + \sum_{k=0}^{\infty} k a_k x^k - \sum_{k=0}^{\infty} a_k x^k \\ &= \sum_{k=0}^{\infty} (k(k-1) + k-1)a_k x^k + \sum_{k=0}^{\infty} k a_k x^{k+1} \\ &= \sum_{k=0}^{\infty} (k^2 - 1)a_k x^k + \sum_{k=1}^{\infty} (k-1)a_{k-1} x^k \\ &= -a_0 + \sum_{k=1}^{\infty} ((k^2 - 1)a_k + (k-1)a_{k-1})x^k, \end{aligned}$$

hence

$$a_0 = 0, \quad (k-1)((k+1)a_k + a_{k-1}) = 0.$$

In conclusion we find

$$\begin{aligned} y(x) &= A_1 \left(x - \frac{x^2}{3} + \frac{x^3}{3 \cdot 4} - \frac{x^4}{3 \cdot 4 \cdot 5} + \cdots \right) \\ &= \frac{2A_1}{x} \left(x - 1 + (1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} - \cdots) \right) = 2A_1 \frac{e^{-x} + x - 1}{x}, \quad x > 0, \end{aligned}$$

i.e., in this case, we find only a one-parameter family of solutions.

7.54 Example. Consider the equation

$$x^3 y'' + y = 0.$$

Assuming $\sum_{k=0}^{\infty} a_k x^k$ is a solution of the equation with a positive radius of convergence we find

$$a_0 + \sum_{k=1}^{\infty} (a_k + (k-1)(k-2)a_{k-1})x^k = 0,$$

hence all a_k must vanish: the unique solution that is representable by a power series with center 0 is the zero solution.

7.55 ¶. All ODEs in this section can be integrated explicitly by writing a *first integral*, i.e., multiplying by y' and obtaining a linear first order ODE for $z(x) := y'(x)$. Of course not all second order equations can be integrated this way.

7.3.4 Generating functions and combinatorics

a. Generating functions

An extremely useful representation of a sequence $\{a_n\}$, $n \geq 0$, that grows less than exponentially is given by the sum of the real or complex power series $\sum_{n=0}^{\infty} a_n z^n$. In fact in this case $\sum_{n=0}^{\infty} a_n z^n$ has positive radius of convergence and the sum $A(z) := \sum_{n=0}^{\infty} a_n z^n$ uniquely identifies its coefficients, see Theorem 7.25. The function $A(z)$ is called the *generating function* of the sequence $\{a_n\}$.

If we restrict ourselves to the set of *bounded sequences* $\{a_n\}$, denoted by $\ell^\infty(\mathbb{C})$, the corresponding series $\sum_{n=0}^{\infty} a_n z^n$ converges to a function defined at least in $\{z \mid |z| < 1\}$. Denoting by \mathcal{C} the set of all maps $a : \{z \mid |z| < 1\} \rightarrow \mathbb{C}$ that are infinitely differentiable in the complex sense, we then establish a map $T : \ell^\infty(\mathbb{C}) \rightarrow \mathcal{C}$, which transforms every bounded sequence $a = \{a_n\}$ into the sum of the corresponding power series

$$T\{a\}(z) := \sum_{n=0}^{\infty} a_n z^n.$$

Since T is injective (see, for example, Theorem 7.25), though not surjective, see Example 6.12, $T\{a\}(z)$ gives a different view of the sequence $\{a_n\}$.

We have

- (i) T is linear, i.e., if $\lambda, \mu \in \mathbb{C}$ and $a = \{a_n\}$, $b = \{b_n\} \in \ell^\infty(\mathbb{C})$, then $\lambda a + \mu b := \{\lambda a_n + \mu b_n\} \in \ell^\infty(\mathbb{C})$ and

$$T\{\lambda a + \mu b\}(z) = \lambda T\{a\}(z) + \mu T\{b\}(z), \quad |z| < 1.$$

- (ii) If $\mathbf{e}_k := \{\underbrace{(0, \dots, 0)}_k, 1, 0, 0, \dots\}$ then $T\{\mathbf{e}_k\}(z) := z^k$.

- (iii) If $a = \{a_n\}$, and

$$b := \{\underbrace{(0, \dots, 0)}_k, a_0, a_1, a_2, \text{dots}\}$$

is the *forward shift* of k places, then

$$T\{b\}(z) = \sum_{n=k}^{\infty} a_n z^{n+k} = z^k T\{a\}(z), \quad |z| < 1.$$

- (iv) If $a = \{a_n\}$, and $b = \{a_{n+k}\}_n$ is the *backward shift* of k places, then

$$T\{b\}(z) = \sum_{n=0}^{\infty} a_{n+k} z^n = \frac{1}{z^k} (T\{a\}(z) - a_0 - a_1 z - a_2 z^2 - \dots - a_{k-1} z^{k-1}).$$

(v) \mathcal{T} transforms the convolution product of sequences into the product of the transformed functions, see Theorem 7.34,

$$\mathcal{T}\{a * b\}(z) = \sum_{n=0}^{\infty} (a * b)_n z^n = \sum_{n=0}^{\infty} a_n z^n \sum_{n=0}^{\infty} b_n z^n = \mathcal{T}\{a\}(z) \mathcal{T}\{b\}(z).$$

7.56 Example. The generating function of the sequence $(1, 1, 1, \dots)$ is

$$\mathcal{T}\{(1, 1, 1, \dots)\}(z) = \sum_{n=0}^{\infty} z^n = \frac{1}{1-z};$$

differentiating, we get

$$\mathcal{T}\{n+1\}(z) = \sum_{n=0}^{\infty} (n+1)z^n = \sum_{n=1}^{\infty} n z^{n-1} = D\left(\frac{1}{1-z}\right) = -\frac{1}{(1-z)^2},$$

while, integrating

$$\mathcal{T}\left\{\frac{1}{n+1}\right\} = -\frac{\log(1-z)}{z}, \quad |z| < 1.$$

7.57 Example. Moreover, if $\mathcal{T}\{a\}(z)$ is the generating function of $a = \{a_n\}$, then

$$\frac{\mathcal{T}\{a\}(z)}{1-z} = \sum_{n=0}^{\infty} z^n \sum_{n=0}^{\infty} a_n z^n = \sum_{n=0}^{\infty} \left(\sum_{k=0}^n a_k\right) z^n = \mathcal{T}\{\bar{a}\}(z)$$

where $\bar{a}_n := \sum_{k=0}^n a_k$. For this reason $1/(1-z)$ is often called the *summing operator*.

7.58 . Let $a = \{a_n\}$ be a sequence that grows at most exponentially fast. We saw that \mathcal{T} is injective, hence it will be possible in principle to reconstruct $\{a_n\}$ from the generating function $\mathcal{T}\{a\}(z)$. Although general formulas are available in the context of the theory of functions of complex variables, it is worth noticing that an explicit formula follows from the Hermite decomposition formula when $\mathcal{T}\{a\}(z)$ is a rational function. Suppose that

$$\sum_{n=0}^{\infty} a_n z^n = \frac{A(z)}{B(z)}$$

in a disc around zero, $A(z)$ and $B(z)$ being coprime polynomials with $\deg A < \deg B$, the roots of B are far from zero, and by Theorem 5.31

$$\frac{A(z)}{B(z)} = \sum_{\alpha \text{ root of } B} \sum_{j=1}^{k_{\alpha}} \frac{\lambda_{\alpha,j}}{(z-\alpha)^j}. \quad (7.24)$$

Since

$$\frac{1}{(z-\alpha)^j} = \frac{(-1)^j}{\alpha^j} \left(1 - \frac{z}{\alpha}\right)^{-j} = \frac{(-1)^j}{\alpha^j} \sum_{n=0}^{\infty} \binom{-j}{n} \frac{z^n}{\alpha^n},$$

on a disc around zero, (7.24) and the principle of identity of power series yields

$$a_n = \sum_{\alpha \text{ root of } B} \frac{1}{\alpha^n} \left(\sum_{j=1}^{k_\alpha} \binom{-j}{n} \frac{(-1)^j \lambda_{\alpha,j}}{\alpha^j} \right) \quad \forall n. \quad (7.25)$$

When all the roots of B are simple, we have $k_\alpha = 1$ and $\lambda_{\alpha,1} = A(\alpha)/B'(\alpha)$. Therefore (7.25) simplifies to

$$a_n = - \sum_{\alpha \text{ root of } B} \frac{A(\alpha)}{\alpha^{n+1} B'(\alpha)}. \quad (7.26)$$

b. Enumerators

Generating functions are particularly useful in combinatorics. In this case it is customary to change slightly the terminology.

7.59 Definition. *The generating function of a sequence $\{a_n\}$ of combinatorial numbers is called the enumerator of $\{a_n\}$.*

7.60 Combinations. The enumerator of the combinations, i.e., of nonordered samples without replacement, in a population of n distinct elements, $\{C_n^k\}_k$,

$$C_n^k := \begin{cases} \binom{n}{k} & \text{if } k \leq n, \\ 0 & \text{if } k > n, \end{cases}$$

is $(1+x)^n$, since by Newton's binomial

$$\sum_{k=0}^{\infty} \binom{n}{k} x^k = \sum_{k=0}^n \binom{n}{k} x^k = (1+x)^n.$$

7.61 Combinations with repetitions. The enumerator of combinations with repetition, or nonordered samples with replacement, from a population of n distinct elements, $C_n^{*k} := \binom{n+k-1}{k}$, $k \geq 0$, is $(1-x)^{-n}$. In fact (see Example 7.40),

$$\sum_{n=0}^{\infty} \binom{n+k-1}{k} x^k = \sum_{n=0}^{\infty} \binom{-n}{k} (-x)^k = \left(\frac{1}{1-x} \right)^n.$$

Enumerators are truly useful since one can code easily several selection rules and constraints. Let us start with some examples.

7.62 Example. From three distinct objects a, b, c , there are three ways to sample one object without replacement, namely a, b or c , three ways to sample two objects without replacement, ab, ac, bc , and only one way to choose three objects, namely abc .

By considering the polynomial $(1+ax)(1+bx)(1+cx)$ and observing that

$$(1+ax)(1+bx)(1+cx) = 1 + (a+b+c)x + (ab+bc+ac)x^2 + (abc)x^3$$

we see that, replacing “or” by $+$ and “and” by \cdot the coefficients of the polynomial enumerate the simultaneous selections of 0, 1, 2, 3 objects.

7.63 Example. The parallelism in the previous example is not casual. Given the four objects a, b, c there are three ways of sampling one object (a, b and c), four ways of choosing two objects (aa, ab, ac, bc), three ways of choosing three objects aab, aac, abc and only one way of choosing four objects ($aabc$). If we encode the population a, b, c by the polynomial

$$P(x) := (1 + ax + a^2x^2)(1 + bx)(1 + cx) \\ = 1 + (a + b + c)x + (a^2 + ab + ac + bc)x^2 + (a^2b + a^2c + abc)x^3 + a^2bcx^4,$$

we see that the coefficient of x^r enumerates the selections of r objects. In particular there are four ways to select two objects, and three ways to select three objects.

7.64 Example. The mechanism is even more general, as we can include constraints on the allowed selections. Still with the population a, b, c , we see that, if we want to count the selections which contain b , it suffices to consider the polynomial

$$(1 + ax + a^2x^2)bx(1 + cx) = bx + (ab + bc)x^2 + (a^2b + abc)x^3 + a^2bcx^4$$

to enumerate the possible selections which contain b .

It is therefore conceivable that we can code the population and the constraints on the element to be selected in a polynomial and leave the job of enumerating the selections to the algebra of polynomials. The previous examples actually suggest how to construct the enumerating polynomial.

Consider a population of N distinct elements a_1, a_2, \dots, a_N but each with multiplicity possibly infinite. For each of the a_i 's consider the power series

$$S_i(x) := \sum_{n=0}^{\infty} \delta_n^i x^n$$

where

$$\delta_n^i := \begin{cases} 1 & \text{if } a_i \text{ may appear } n \text{ times in the selection,} \\ 0 & \text{if } a_i \text{ is not allowed to appear } n \text{ times in the selection.} \end{cases}$$

The product of these series, one for each distinguishable element of the population, all converging in $] -1, 1[$, $S_1(x)S_2(x) \cdots S_N(x)$, is the enumerator of the drawing.

7.65 Example. In the case of combinations without repetitions of a population of n distinct elements, that is of unordered samples without repetitions, each element may appear at most once. Thus its enumerator is $(1 + x)$ and the enumerator of the combinations without repetitions is

$$\underbrace{(1 + x)(1 + x) \cdots (1 + x)}_{n \text{ times}} = (1 + x)^n.$$

7.66 Example. In the case of sampling with replacement, the population has n distinct elements, but each can occur with arbitrary multiplicity. The enumerator of each element is then

$$1 + x + x^2 + \cdots = \frac{1}{1 - x}$$

and the enumerator of unordered samples with repetitions is

$$\underbrace{\left(\sum_{k=0}^{\infty} x^k\right) \cdot \left(\sum_{k=0}^{\infty} x^k\right) \cdots \left(\sum_{k=0}^{\infty} x^k\right)}_{n\text{-times}} = \left(\sum_{k=0}^{\infty} x^k\right)^n = \left(\frac{1}{1-x}\right)^n.$$

7.67 ¶. Show that $\sum_{r=0}^n (-1)^r \binom{n}{r} = 0$, i.e., the ways of choosing an even or an odd number of objects is equal, and equal to 2^{n-1} .

7.68 ¶. Show that $\sum_{k=0}^n \binom{n}{k}^2 = \binom{2n}{n}$. [Hint: $(1+z)^n(1+z)^n = (1+z)^{2n}$.]

7.69 ¶. Prove Vandermonde's formula using the identity

$$(1+z)^N = (1+z)^K (1+z)^{N-K}.$$

c. Exponential enumerators

For sequences $\{a_n\}$, and especially for sequences which grow faster than exponentially, it is worth computing the enumerator of a rescaled sequence. For instance the *exponential enumerator* of the sequence $\{a_n\}$ is the sum of the power series

$$\sum_{n=0}^{\infty} a_n \frac{z^n}{n!}.$$

7.70 Arrangements without repetitions. The enumerator of the ordered samples, or arrangements, without repetitions of n distinct objects $\{D_n^k\}_k$,

$$D_n^k := \begin{cases} \frac{n!}{n-k!} & \text{if } k \leq n, \\ 0 & \text{if } k > n \end{cases}$$

of n distinct objects is

$$\sum_{k=0}^n D_n^k x^k = 1 + \frac{n!}{(n-1)!}x + \frac{n!}{(n-2)!}x^2 + \cdots + n!x^n,$$

which unfortunately has no simple closed form. However the *exponential enumerator* of the same sequence D_n^k is

$$\sum_{k=0}^n \frac{D_n^k}{k!} x^k = (1+x)^n.$$

7.71 Arrangements with repetitions. The exponential enumerator of the ordered k -samples of n with repetitions of n distinct objects, $\{D_n^{*k}\}$, $D_n^{*k} = n^k$, is

$$\sum_{k=0}^{\infty} \frac{n^k}{k!} x^k = e^{nx}.$$

As for nonordered samples, one can build easily from the population and the rules of selection the *exponential generator* for nonordered samples, leaving the computation to the algebra of the power series.

Suppose the population is made up of N distinct elements, a_1, a_2, \dots, a_N , each with infinite multiplicity. For each a_i , consider the power series

$$S_i(x) := \sum_{n=0}^{\infty} \delta_n^i \frac{x^n}{n!}$$

where

$$\delta_n^i := \begin{cases} 1 & \text{if } a_i \text{ may appear } n \text{ times in the selection,} \\ 0 & \text{if } a_i \text{ may not appear } n \text{ times in the selection.} \end{cases}$$

Then the exponential enumerator of the ordered samples with repetitions is $S_1(x)S_2(x) \cdots S_N(x)$.

7.72 ¶. Check that the previous rule yields the right result in the case of permutations with or without repetitions.

7.73 ¶. Show that the exponential enumerator of the permutations of

- p identical objects is $\frac{z^p}{p!}$,
- two objects of one type and three of another type is

$$\left(1 + x + \frac{x^2}{2!}\right) \left(1 + x + \frac{x^2}{2!} + \frac{x^3}{3!}\right).$$

d. A few location problems

The exponential enumerator is particularly useful when locating distinct objects into cells.

7.74 Distributions onto distinct cells and surjective maps. As we have already stated, locating k different objects in n different cells is equivalent to fixing a map from $X := \{1, \dots, k\}$ into $Y := \{1, \dots, n\}$. Thus, the number of ways of placing k distinct objects in n distinct cells with no cell left empty is equal to the number S_n^k of surjective maps from X into Y . By Proposition 3.38,

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

We give an alternate simple proof based on the use of the exponential enumerator. Since we have n cells and each cell may contain an arbitrary number of objects larger than 1, the exponential enumerator for each cell is

$$x + x^2 + x^3 + \cdots = \sum_{k=1}^{\infty} \frac{x^k}{k!} = e^x - 1,$$

hence the exponential enumerator of the distribution in n cells is

$$\begin{aligned}
 (e^x - 1)^n &= \sum_{j=0}^n \binom{n}{j} (-1)^j e^{(n-j)x} \\
 &= \sum_{j=0}^n \binom{n}{j} (-1)^j \sum_{k=0}^{\infty} (n-j)^k \frac{x^k}{k!} \\
 &= \sum_{k=0}^{\infty} \left(\sum_{j=0}^n \binom{n}{j} (-1)^j (n-j)^k \right) \frac{x^k}{k!}.
 \end{aligned} \tag{7.27}$$

The number of k -permutations of the n cells being the coefficient of $x^k/k!$, we find again the value of S_n^k . The numbers

$$S(k, n) := \frac{1}{n!} S_n^k = \frac{1}{n!} \sum_{j=0}^n \binom{n}{j} (-1)^j (n-j)^k$$

are called *Stirling numbers of second kind* and (7.27) can be rewritten as

$$\frac{(e^x - 1)^n}{n!} = \sum_{k=0}^{\infty} S(k, n) \frac{x^k}{k!}. \tag{7.28}$$

7.75 Distributions into indistinct cells. Since there are $n!$ ways of distinguishing n objects *Stirling number* $S(k, n)$ is the number of ways of placing k distinct objects into n nondistinct cells, all containing at least one element.

We also saw that there are n^k ways of placing k distinct objects into n distinct cells, when empty cells are allowed. However, the number of ways of distributing k distinct objects in n nondistinct cells with empty cells allowed is not $n^k/n!$. It is

$$S(k, 1) + S(k, 2) + \cdots + S(k, n) \quad \text{for } k > n$$

and

$$S(k, 1) + S(k, 2) + \cdots + S(k, k) \quad \text{for } k \leq n,$$

i.e., in both cases $\sum_{j=0}^{\min(n, k)} S(k, j)$. In fact, the ways of distributing k objects in n nondistinct cells with empty cells allowed equals the ways of distributing the k objects so that one cell is not empty, or two cells are not empty, etc.

If $n \geq k$ the number $\sum_{j=0}^k S(k, j)$ of distributions of k objects into n -distinct cells has a closed form. In fact, since $S(k, n) = 0$ for $n > k$, we have

$$\sum_{j=0}^k S(k, j) = \sum_{j=0}^{\infty} S(k, j)$$

consequently

$$\begin{aligned}\sum_{j=0}^n S(k, j) &= \sum_{j=0}^{\infty} \frac{1}{j!} \sum_{n=0}^j (-1)^n \binom{j}{n} (j-n)^k = \sum_{j=0}^{\infty} \sum_{n=0}^j \frac{(-1)^n}{n!} \frac{(j-n)^k}{(j-n)!} \\ &= \left(\sum_{j=0}^{\infty} \frac{(-1)^j}{j!} \right) \left(\sum_{j=0}^{\infty} \frac{j^k}{j!} \right) = \frac{1}{e} \sum_{j=0}^{\infty} \frac{j^k}{j!}.\end{aligned}\tag{7.29}$$

Moreover, from (7.28)

$$\begin{aligned}\sum_{k=0}^{\infty} \sum_{j=0}^k S(k, j) \frac{x^k}{k!} &= \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} S(k, j) \frac{x^k}{k!} = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} S(k, j) \frac{x^k}{k!} \\ &= \sum_{j=0}^{\infty} \frac{(e^x - 1)^j}{j!} = e^{e^x - 1}.\end{aligned}\tag{7.30}$$

e. Partitions of a set

The exponential enumerator is very useful also when dealing with partitioning. Let $X_k := \{1, 2, 3, \dots, k\}$ be a set with k elements. A *partition* of X_k is a decomposition of X_k into a finite union of disjoint subsets C_1, C_2, \dots, C_p . We denote by $P := \{C_1, C_2, \dots, C_p\}$ a partition and by $P(X_k)$ the family of partitions of X_k . Two partitions $\{C_1, C_2, \dots, C_p\}$ and $\{D_1, D_2, \dots, D_q\}$ are different if $p \neq q$ or, if $p = q$ and for any permutation σ of the indices we can find i such that $C_i \neq D_{\sigma(i)}$. The number of partitions of X_k , called the k -th *Bell number*, equals the number of distributions of k distinct objects into r cells allowing empty cells, $r \geq k$, hence by (7.29)

$$|P(X_k)| = \sum_{j=0}^{\infty} S(k, j) = \frac{1}{e} \sum_{j=0}^{\infty} \frac{j^k}{j!}.$$

We can also find such a number by means of the exponential enumerators. For that we first state

7.76 Proposition. Let $u(x) := \sum_{k=1}^{\infty} a_k \frac{x^k}{k!}$ be the exponential enumerator of the sequence $\{a_n\}$, $a_0 = 0$. Then

$$e^{u(x)} = \sum_{k=0}^{\infty} \frac{A_k}{k!} x^k \quad \text{where} \quad A_k = \sum_{P \in P(X_k)} \prod_{C \in P} a_{|C|},$$

$|C|$ being the cardinality of C .

Proof. In fact

$$\begin{aligned}
e^{u(x)} &:= \sum_{j=0}^{\infty} \frac{u^j(x)}{j!} = \sum_{j=0}^{\infty} \frac{1}{j!} \left(\sum_{k=1}^{\infty} a_k \frac{x^k}{k!} \right)^j \\
&= \sum_{j=0}^{\infty} \frac{1}{j!} \left(\sum_{k_1, k_2, \dots, k_j=1, \infty} a_{k_1} a_{k_2} \cdots a_{k_j} \frac{x^{k_1+k_2+\dots+k_j}}{k_1! k_2! \cdots k_j!} \right) = 1 + \sum_{r=1}^{\infty} A_r \frac{x^r}{r!}
\end{aligned}$$

where

$$A_r := \sum_{j=1}^r \sum_{\substack{k_1+k_2+\dots+k_j=r \\ k_i \geq 1}} a_{k_1} a_{k_2} \cdots a_{k_j} \frac{r!}{j! k_1! k_2! \cdots k_j!}. \quad (7.31)$$

We may interpret k_1, k_2, \dots, k_j as the cardinalities of a partition $P = (C_1, C_2, \dots, C_j)$ of $\{1, 2, \dots, r\}$ and let P_{j, k_1, \dots, k_j} be the set of partitions with j subsets of cardinality k_1, \dots, k_j . Since the number of partitions of $\{1, 2, \dots, r\}$ in j subsets, C_1, C_2, \dots, C_j , with cardinality k_1, \dots, k_j is

$$\frac{r!}{k_1! k_2! \cdots k_j!}$$

(see, for example, 3.51), we conclude from (7.31) that

$$A_r = \sum_{j=1}^r \sum_{P \in \mathcal{P}_{j, k_1, \dots, k_j}} \left(\prod_{C \in P} a_{|C|} \right) = \sum_{P \in P(X_r)} \prod_{C \in P} a_{|C|}.$$

□

To compute the number of partitions of X_n , we now choose $a_k = 1$, $k \geq 1$, hence $u(x) = \sum_{k=1}^{\infty} a_k x^k / k! = e^x - 1$ and Proposition 7.76 yields

$$e^{e^x - 1} = \sum_{n=0}^{\infty} |P(X_n)| \frac{x^n}{n!}.$$

On the other hand by (7.30)

$$e^{e^x - 1} = \sum_{k=0}^{\infty} \sum_{j=0}^{\infty} S(k, j) \frac{x^k}{k!},$$

therefore $|P(X_n)| = \sum_{j=0}^{\infty} S(n, j)$, and, by (7.29), $|P(X_n)| = \frac{1}{e} \sum_{k=0}^{\infty} \frac{k^n}{k!}$.

7.77 ¶¶. If in Proposition 7.76 we choose $a_k = \#$ of trees with k vertices and $u(x)$ is the corresponding exponential enumerator, then the coefficients A_k in

$$e^{u(x)} = \sum_{k=0}^{\infty} A_k \frac{x^k}{k!}$$

represent the forests of trees with k vertices.

7.78 Partition of integers. In how many different ways can we decompose an integer as a sum of integers? For example the number 4 can be decomposed as 4, 3 + 1, 2 + 2, 2 + 1 + 1, 1 + 1 + 1 + 1. This was one of the problems discussed by Leonhard Euler (1707–1783) in terms of generating functions. Clearly, a partition of n is equivalent to a way of distributing n nondistinct cells with empty cells allowed. It is not difficult to realize that the generating function of the sequence $\{p(n) \mid p(n) := \# \text{ partitions of } n\}$ is given by

$$\begin{aligned} \sum_{k=0}^{\infty} p_k x^k &= (1 + x + x^2 + \cdots + x^r + \cdots) \\ &\quad \cdot (1 + x^2 + x^4 + \cdots) \\ &\quad \cdot (1 + x^3 + x^6 + \cdots) \\ &\quad \cdots \\ &= \prod_{k=1}^{\infty} \frac{1}{1 - x^k}. \end{aligned} \quad (7.32)$$

Similarly, observing that

$$1 + x + x^2 + x^3 + \cdots = \frac{1}{1 - x} = (1 + x)(1 + x^2)(1 + x^4) \cdots (1 + x^{2^r}),$$

we infer that any integer can be expressed as the sum of a selection of nonnegative integral powers of 2 (without repetitions) exactly in one way, i.e., every decimal can be represented uniquely as a binary alignment.

Ut non-finitam Seriem finita cöercet,
 Summula, & in nullo limite limes adest:
 Sic modico immensi vestigia Numinis haerent
 Corpore, & angusto limite limes abest.
 Cernere in immenso parvum, dic, quanta voluptas!
 In parvo immensum cernere, quanta, Deum!
Jacob Bernoulli⁵

⁵ Even as the finite encloses an infinite series

And in the unlimited limits appear,
 So the soul of immensity dwells in minutia
 And in narrowest limits inhere.

What joy to discern the minute in infinity!

The vast to perceive in the small, what divinity!

From Jacob Bernoulli, *Tractatus de Seriebus infinitis Earumque Summa Finita et Usu in Quadraturis Spatorum & Rectificationibus Curvarum*, in *Ars Conjectandi* (Translation by Helen M. Walker, from *A Source Book in Mathematics* by D. E. Smith, 1929).

7.4 Further Applications

In this section we illustrate some applications, which go back to Euler and Johann Bernoulli (1667–1748) and are quite relevant in several contexts.

7.4.1 Euler–MacLaurin summation formula

In this section we illustrate a general method of approximating sums found by Euler and later rediscovered by Colin MacLaurin (1698–1746). As a consequence we find the asymptotic development of the factorial $n!$ and of the partial sums of the harmonic series, $H_n := \sum_{k=1}^n \frac{1}{k}$.

a. Bernoulli numbers

As we shall see, the Taylor series of the function

$$g(z) := \begin{cases} \frac{z}{e^z - 1}, & z \neq 0, \\ 1, & z = 0 \end{cases}$$

with center in the origin plays an important role. From the theory of complex functions one infers that g has a Taylor expansion with radius of convergence 2π , so we can write for $|z| < 2\pi$,

$$\frac{z}{e^z - 1} = \sum_{j=0}^{\infty} B_j \frac{z^j}{j!}. \quad (7.33)$$

The numbers $\{B_j\}$ are called the *Bernoulli numbers*. From

$$\begin{aligned} 0 &= (e^z - 1) \sum_{n=0}^{\infty} B_n \frac{z^n}{n!} - z \\ &= \sum_{n=0}^{\infty} \sum_{j=0}^{\infty} B_j \frac{z^j}{j!} \frac{z^{n+1-j}}{(n+1-j)!} - z \\ &= \sum_{n=1}^{\infty} \sum_{j=0}^n \binom{n+1}{j} B_j \frac{z^{n+1}}{(n+1)!}, \end{aligned}$$

we see that they are characterized by the implicit recurrence relation

$$\begin{cases} B_0 := 1, \\ \sum_{j=0}^n \binom{n+1}{j} B_j = 0 \quad \forall n \geq 1, \end{cases} \quad (7.34)$$

from which we can easily compute a few values of B_n :

$$B_0 = 1, \quad B_1 = -\frac{1}{2}, \quad B_2 = \frac{1}{6}, \quad B_3 = 0, \quad B_4 = -\frac{1}{30}, \quad B_5 = 0.$$

7.79 . Convergence and equality in (7.34), and other consequences, can be inferred starting from Euler's formula for $\cot z$, compare (6.30),

$$z \cot z = 1 - 2 \sum_{k=1}^{\infty} \frac{z^2}{k^2 \pi^2 - z^2}, \quad |z| < \pi. \quad (7.35)$$

On account of the Weierstrass double series theorem, we infer that

$$\begin{aligned} z \cot z &= 1 - 2 \sum_{k=1}^{\infty} \frac{z^2}{k^2 \pi^2 - z^2} \\ &= 1 - 2 \sum_{k=1}^{\infty} \frac{z^2}{k^2 \pi^2} \frac{1}{1 - \frac{z^2}{k^2 \pi^2}} \\ &= 1 - 2 \sum_{k=1}^{\infty} \sum_{j=0}^{\infty} \left(\frac{z^2}{k^2 \pi^2} \right)^{j+1} \\ &= 1 - 2 \sum_{j=1}^{\infty} \alpha_{2j} z^{2j} \end{aligned} \quad (7.36)$$

where

$$\alpha_{2j} := \frac{1}{\pi^{2j}} \left(\sum_{k=1}^{\infty} \frac{1}{k^{2j}} \right).$$

The equality (7.36) holds on $|z| < \pi$. On the other hand

$$\frac{z}{e^z - 1} + \frac{z}{2} = \frac{z}{2} \frac{e^z + 1}{e^z - 1} = \frac{z}{2} \frac{e^{z/2} + e^{-z/2}}{e^{z/2} - e^{-z/2}} \quad (7.37)$$

$$= \frac{z}{2} \frac{\cosh(\frac{z}{2})}{\sinh(\frac{z}{2})} = \frac{z}{2} \coth\left(\frac{z}{2}\right) = w \cot w, \quad (7.38)$$

where $2i w := z$. Equating (7.36) and (7.37), we then infer

$$\frac{z}{e^z - 1} = 1 - \frac{z}{2} - 2 \sum_{j=1}^{\infty} \alpha_{2j} w^{2j} = 1 - \frac{z}{2} - 2 \sum_{j=1}^{\infty} \frac{(-1)^j}{4^j} \alpha_{2j} z^{2j}$$

near zero, hence $z/(e^z - 1)$ has a Taylor expansion centered at zero,

$$\frac{z}{e^z - 1} = \sum_{j=0}^{\infty} B_j \frac{z^j}{j!}$$

where $B_0 = 1$, $B_1 = 1/2$, and, for $j \geq 1$, $B_{2j+1} = 0$ and

$$\frac{B_{2j}}{(2j)!} := \frac{(-1)^{j-1}}{2^{2j-1} \pi^{2j}} \sum_{k=1}^{\infty} \frac{1}{k^{2j}}. \quad (7.39)$$

In particular

$$\frac{|B_{2n}|}{(2n)!} = \frac{2}{(2\pi)^{2n}} \sum_{k=1}^{\infty} \frac{1}{k^{2n}} \leq 2 \sum_{k=1}^{\infty} \frac{1}{k^2} \frac{1}{(2\pi)^{2n}}, \quad (7.40)$$

from which we infer that

$$S(z) := \sum_{j=1}^{\infty} B_j \frac{z^j}{j!}$$

has radius of convergence at least 2π . Since $S(z)$ and $z/(e^z - 1)$ have both complex derivatives on $|z| < 2\pi$, we conclude that

$$\frac{z}{e^z - 1} = \sum_{j=0}^{\infty} B_j \frac{z^j}{j!}, \quad |z| < 2\pi.$$

Notice that (7.40) yields

$$\sum_{k=1}^{\infty} \frac{1}{k^{2j}} = (-1)^{j-1} \frac{2^{2j-1} \pi^{2j} B_{2j}}{(2j)!}, \quad j \geq 1;$$

in particular,

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}, \quad \sum_{k=1}^{\infty} \frac{1}{k^4} = \frac{\pi^4}{90}.$$

b. Bernoulli polynomials

Bernoulli polynomials are defined by

$$B_n(x) := \sum_{k=0}^n \binom{n}{k} B_k x^{n-k}, \quad x \in \mathbb{R}. \quad (7.41)$$

It is not difficult to show that the exponential enumerator of $\{B_n(x)\}$ is $te^{xt}/(e^t - 1)$, i.e.,

$$\frac{te^{xt}}{e^t - 1} = \sum_{k=0}^{\infty} B_k(x) \frac{t^k}{k!}, \quad |t| < 2\pi. \quad (7.42)$$

They satisfy the relations

$$B_n(x+1) - B_n(x) = nx^{n-1}, \quad (7.43)$$

$$DB_n(x) = nB_{n-1}(x). \quad (7.44)$$

In particular

$$\begin{aligned} B_0(x) = B_0 = 1, & \quad B_n(1) = B_n(0) = B_n \quad \forall n \geq 1, \\ \int_0^1 B_n(x) dx = 0, & \quad \int_0^x B_{n-1}(t) dt = \frac{B_n(x) - B_n}{n}. \end{aligned} \quad (7.45)$$

To prove (7.43) we compute

$$\sum_{n=0}^{\infty} \left(B_n(x+1) - B_n(x) \right) \frac{t^n}{n!} = te^{xt} = \sum_{n=1}^{\infty} x^{n-1} \frac{t^n}{(n-1)!},$$

then it remains to equate the coefficients of t^n for all n . To prove (7.44) it suffices to differentiate (7.41), while the rest is then trivial.

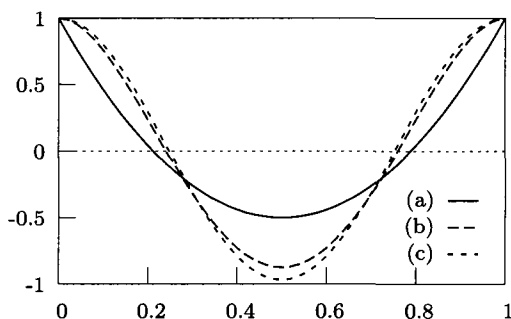


Figure 7.3. The normalized Bernoulli polynomials $B_n(x)/B_n$, respectively (a) $n = 2$, (b) $n = 4$ and (c) $n = 6$.

7.80 . From (7.43), we infer

$$B_{m+1}(n) - B_{m+1}(1) = \sum_{j=1}^{m-1} (B_{m+1}(j+1) - B_{m+1}(j)) = (m+1) \sum_{j=1}^n j^m,$$

that is,

$$\sum_{k=0}^{\infty} k^m = \frac{1}{m+1} (B_{m+1}(n) - B_{m+1}(1)).$$

7.81 . Using the properties of Bernoulli polynomials, in particular (7.44) and (7.45), one proves inductively on n (and we leave it to the reader) that the only possible minimum and maximum points for $B_{2m}(x)$, $x \in \mathbb{R}$, are 0, 1 and $1/2$. From

$$\sum_{n=0}^{\infty} B_m(1/2) \frac{x^n}{n!} = \frac{x e^{x/2}}{e^x - 1} = \frac{x}{e^{x/2} - 1} - \frac{x}{e^x - 1}$$

we compute

$$B_{2m}(1/2) = (2^{1-2m} - 1)B_{2m}$$

consequently

$$\|B_{2m}\|_{\infty, [0,1]} = \sup_{t \in [0,1]} |B_{2m}(t)| \leq |B_{2m}|. \quad (7.46)$$

c. Euler–MacLaurin formula and Stirling’s approximation

7.82 Theorem. Given nonnegative integers m, p, q , $m \geq 1$, $p \leq q$, and a smooth function f , we have

$$\sum_{k=p}^{q-1} f(k) = \int_p^q f(x) dx + \sum_{k=1}^m \frac{B_k}{k!} D^{k-1} f(x) \Big|_p^q + R_m \quad (7.47)$$

where

$$R_m := (-1)^{m+1} \int_p^q \frac{B_m(x - [x])}{m!} D^m f(x) dx.$$

Proof. To prove it, we first observe that it suffices to prove it in the case $p = 0$ and $q = 1$, because we can then replace f by $f(x + h)$ for any integer h getting

$$\begin{aligned} f(h) &= \int_h^{h+1} f(x) dx + \sum_{k=1}^m \frac{B_k}{k!} D^{k-1} f(x) \Big|_h^{h+1} \\ &\quad - (-1)^m \int_h^{h+1} \frac{B_m(x - [x])}{m!} D^m f(x) dx; \end{aligned}$$

summing in h on the range $p \leq h < q$, we then get (7.47), since intermediate terms telescope nicely. The proof when $p = 0$ and $q = 1$, i.e., of

$$\begin{aligned} f(0) &= \int_0^1 f(x) dx + \sum_{k=1}^m \frac{B_k}{k!} D^{k-1} f(x) \Big|_0^1 + R_m, \\ R_m &:= (-1)^{m+1} \int_0^1 \frac{B_m(x)}{m!} D^m f(x) dx \end{aligned}$$

is by induction on m . For $m = 1$ it amounts to proving

$$f(0) = \int_0^1 f(x) dx - \frac{1}{2} (f(1) - f(0)) + \int_0^1 (x - 1/2) f'(x) dx$$

which is just

$$\frac{f(1) + f(0)}{2} = \int_0^1 D((x - 1/2)f(x)) dx = \int_0^1 f(x) dx + \int_0^1 (x - 1/2) f'(x) dx.$$

To pass from $m - 1$ to m , $m > 1$, we need to show that

$$R_{m-1} := \frac{B_m}{m!} D^{m-1} f(x) \Big|_0^1 + R_m$$

which reduces to

$$(-1)^m B_m D^{m-1} f(x) \Big|_0^1 = m \int_0^1 B_{m-1}(x) D^{m-1} f(x) dx + \int_0^1 B_m(x) D^m f(x) dx.$$

As previously, taking into account (7.44), integrating by parts we see that this holds if and only if

$$(-1)^m B_m D^{m-1} f(x) \Big|_0^1 = B_m(x) D^{m-1} f(x) \Big|_0^1,$$

i.e., if and only if

$$(-1)^m B_m = B_m(1) = B_m(0) \quad \forall m > 1,$$

that we know to hold since $B_m(1) = B_m(0) = B_m$, and $B_m = 0$ for m odd. \square

On account of (7.46) and (7.40), we can easily evaluate the remainder and rewrite the Euler–MacLaurin formula as

$$\sum_{k=p}^{q-1} f(k) = \int_p^q f(x) dx - \frac{1}{2} f(x) \Big|_p^q + \sum_{k=1}^m \frac{B_{2k}}{(2k)!} D^{2k-1} f(x) \Big|_p^q + R_m, \quad (7.48)$$

with

$$|R_{2m}| \leq \frac{|B_{2m}|}{(2m)!} \int_p^q |D^{2m} f(x)| dx. \quad (7.49)$$

If $D^{2m} f(x) \geq 0$ in $[p, q]$, $D^{2m-1} f(x)$ is increasing and the integral $\int_p^q |D^{2m} f(x)| dx$ is just $D^{2m-1} f(x) \Big|_p^q$, therefore, using (7.46), we can estimate the remainder by

$$\begin{aligned} |R_{2m}| &\leq \int_p^q \frac{|B_{2m}(x - [x])|}{(2m)!} |D^{2m} f(x)| dx \\ &\leq \frac{\|B_{2m}(x)\|_{\infty, [0,1]}}{(2m)!} \int_p^q |D^{2m} f(x)| dx \\ &\leq \frac{B_{2m}}{(2m)!} D^{2m-1} f(x) \Big|_p^q. \end{aligned} \quad (7.50)$$

7.83 . An interesting application of the Euler–MacLaurin formula is to the study of the asymptotic development of $\sum_{k=1}^n f(k)$ when $n \rightarrow \infty$. The structure of the Euler–MacLaurin formula is

$$\sum_{k=1}^{n-1} f(k) = F(n) - F(1) + \sum_{k=1}^m (T_k(n) - T_k(1)) + R_m(n)$$

where

$$R_m(n) := \int_0^n \frac{B_m(x - [x])}{m!} D^m f(x) dx,$$

etc. Assuming $D^m f(x) = O(x^{c-m})$ as $x \rightarrow \infty$ for large m , it is not difficult to see that $R_m(n)$ is not small for large n , but has only a small tail, i.e.,

$$\begin{cases} R_m(n) = R_m(\infty) + \widehat{R}_m(n), \\ \widehat{R}_m(n) = (-1)^{m+1} \int_n^\infty \frac{B_m(x - [x])}{m!} D^m f(x) dx = O(n^{c+1-m}). \end{cases}$$

Therefore we can conclude that for a suitable constant C we have

$$\sum_{k=1}^{n-1} f(k) = F(n) + C + \sum_{k=1}^m T_k(n) + \widehat{R}_m(n).$$

7.84 Example (Harmonic series). We apply 7.83 to $f(x) := 1/x$. Since

$$D^k f(x) = (-1)^k k! / x^{k+1},$$

we deduce

$$\sum_{k=1}^{n-1} \frac{1}{k} = \log n + C + \frac{B_1}{n} - \sum_{k=1}^m \frac{B_{2k}}{2kn^{2k}} + R'_m(n)$$

for some constant C . Since $D^{2m}f(x) \geq 0 \forall m$, we can estimate the rest by

$$|R'_m(n)| \leq \frac{B_{2m+2}}{(2m+2)n^{2m+2}};$$

adding $1/n$ and observing that C is the Euler–Mascheroni constant γ , see Example 6.26, we conclude

$$H_n := \sum_{k=1}^n \frac{1}{k} = \log n + \gamma + \frac{1}{2n} - \sum_{k=1}^m \frac{B_{2k}}{2kn^{2k}} + \theta_{m,n} \frac{B_{2m+2}}{(2m+2)n^{2m+2}}$$

for some $\theta_{m,n}$ with $|\theta_{m,n}| \leq 1$.

7.85 Example (Stirling approximation). Similarly to Example 7.84, on account of Stirling's formula in Example 2.67, we can state

$$\begin{aligned} \log n! &= n \log n - n + \frac{1}{2} \log n + \log \sqrt{2\pi} + \sum_{k=1}^m \frac{B_{2k}}{2k(2k-1)n^{2k-1}} \\ &\quad + \theta_{m,n} \frac{B_{2m+2}}{(2m+2)(2m+1)n^{2m+1}} \end{aligned}$$

and $|\theta_{m,n}| \leq 1$.

7.4.2 Euler Γ function

The gamma function, $\Gamma(x)$, defined by Euler in 1729, is surely one of the most important *special functions*, as it unexpectedly appears in many topics in analysis.

a. Definition and characterizations

For $0 < x < \infty$, $\Gamma(x)$ is defined as

$$\Gamma(x) := \int_0^\infty t^{x-1} e^{-t} dt.$$

7.86 Proposition. $\Gamma(x) \in \mathbb{R}_+$ and, for all $x \in]0, \infty[$,

- $\Gamma(x+1) = x\Gamma(x)$,
- $\Gamma(1) = 1$, $\Gamma(n+1) = n!$,
- $\log \Gamma(x)$, $x \in]0, \infty[$ is convex on $]0, \infty[$, in particular Γ is a continuous function.

Proof. (i) follows easily integrating by parts. Clearly $\Gamma(1) = 1$, thus (ii) follows from (i) by induction. Applying Hölder's inequality (see [GM1]), one easily obtains

$$\Gamma\left(\frac{x}{p} + \frac{y}{q}\right) \leq \Gamma(x)^{1/p} \Gamma(y)^{1/q}, \quad \frac{1}{p} + \frac{1}{q} = 1,$$

that is equivalent to (iii). □

Actually these three properties characterize $\Gamma(x)$ completely.

7.87 Theorem. Let $f :]0, \infty[\rightarrow]0, \infty[$ be a function such that

- $f(x+1) = xf(x)$,
- $f(1) = 1$,
- $\log f$ is convex.

Then $f(x) = \Gamma(x) \forall x > 0$.

Proof. It suffices to show that (i) (ii) (iii) uniquely determines $f(x)$ for all $x > 0$ and actually because of (i), for all $x \in]0, 1[$. Set $\varphi := \log f$. Then

$$\varphi(x+1) = \varphi(x) + \log x, \quad \varphi(1) = 0 \quad \text{and} \quad \varphi \text{ is convex.} \quad (7.51)$$

By induction we see that $\varphi(n+1) = \log n!$ for all integers $n \geq 1$. Since φ is convex (see, e.g., [GM1]), we have for $0 < x < 1$,

$$\begin{aligned} \log n &= \frac{\varphi(n+1) - \varphi(n)}{1} \leq \frac{\varphi(n+1+x) - \varphi(n+1)}{x} \\ &\leq \frac{\varphi(n+2) - \varphi(n+1)}{1} = \log(n+1), \end{aligned} \quad (7.52)$$

while iterating the first of (7.51),

$$\varphi(n+1+x) = \varphi(x) + \log[x(x+1) \cdots (n+n)].$$

Subtracting $\log n$ in (7.52), we then get

$$0 \leq \varphi(x) - \log \frac{n! n^x}{x(x+1) \cdots (x+n)} \leq x \log \left(1 + \frac{1}{n}\right).$$

Since the last term tends to zero as $n \rightarrow \infty$, $\varphi(x)$ is uniquely determined. \square

7.88 Gauss's formula. In the proof of Theorem 7.87 we have in fact proved that

$$\Gamma(x) = \lim_{n \rightarrow \infty} \frac{n! n^x}{x(x+1) \cdots (x+n)},$$

or, equivalently

$$\lim_{n \rightarrow \infty} \frac{\Gamma(x+n)}{n^x \Gamma(n)} = 1.$$

Actually one can prove that the previous formula is a characteristic for gamma. We have the following.

7.89 Theorem. Let $F :]0, \infty[\rightarrow]0, \infty[$ be a function such that

- (i) $F(x+1) = xF(x)$,
- (ii) $F(1) = 1$,
- (iii) $\lim_{n \rightarrow \infty} \frac{F(x+n)}{n^x F(n)} = 1$.

Then $F(x) = \Gamma(x)$.

Proof. In fact,

$$F(n) = (n-1)! \quad \text{and} \quad F(x+n) = x(x+1)\cdots(x+n-1)F(x),$$

therefore (iii) yields

$$1 = \lim_{n \rightarrow \infty} \frac{F(x+n)}{n! n^{x-1}} = F(x) \lim_{n \rightarrow \infty} \frac{x(x+1)\cdots(x+n-1)}{n! n^{x-1}} = \frac{F(x)}{\Gamma(x)}.$$

□

We can also express $\Gamma(x)$ as an infinite product: this is the original definition of Euler.

7.90 Proposition. *We have*

$$\frac{1}{\Gamma(x)} = e^{\gamma x} x \prod_{n=1}^{\infty} \left(1 + \frac{x}{n}\right) e^{-x/n},$$

where γ is the Euler–Mascheroni constant.

Proof. Write $g(x)$ for the inverse of $\Gamma(x)$ in the right-hand side. Taking the logarithm we see that $g(1) = 1$ and

$$\begin{aligned} \frac{1}{g(x)} &= \lim_{n \rightarrow \infty} \exp \left(x \left(1 + \frac{1}{2} + \cdots + \frac{1}{n} - \log n \right) \right) x \prod_{j=1}^n \left(1 + \frac{x}{j} \right) e^{-x/j} \\ &= \lim_{n \rightarrow \infty} \exp(-x \log n) x \prod_{j=1}^n \left(1 + \frac{x}{j} \right) \\ &= \lim_{n \rightarrow \infty} \frac{x \left(1 + \frac{x}{1} \right) \left(1 + \frac{x}{2} \right) \cdots \left(1 + \frac{x}{n} \right)}{n^x} \end{aligned}$$

i.e.,

$$g(x) = \lim_{n \rightarrow \infty} \frac{n! n^x}{x(x+1)\cdots(x+n)}.$$

□

b. Functional relations

7.91 Beta function. The function

$$B(x, y) := \int_0^1 t^{x-1} (1-t)^{y-1} dt, \quad x, y > 0$$

is called the *beta function*. It is related to the gamma function by

Proposition. *We have*

$$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}; \quad (7.53)$$

In particular, since $\gamma(n+1) = n!$,

$$B(m+1, n+1) = \frac{1}{\binom{n+m+1}{m}}$$

for all $n, m \in \mathbb{N}$.

Proof. Set

$$f(x) := \frac{\Gamma(x+y)}{\Gamma(y)} B(x, y).$$

We have

$$f(1) := \frac{\Gamma(1+y)}{\Gamma(y)} B(1, y),$$

since $B(1, y) = 1/y$. $\log f$ is convex, since $x \rightarrow B(x, y)$ is convex: this can be proved as in Proposition 7.86. Finally,

$$\begin{aligned} f(x+1) &= \frac{\Gamma(x+y+1)}{\Gamma(y)} B(x+1, y) \\ &= (x+y) \frac{\Gamma(x+y)}{\Gamma(y)} B(x+1, y) = x f(x), \end{aligned}$$

since $B(x+1, y) = \frac{x}{x+y} B(x, y)$ as it is easily seen by performing an integration by parts. Theorem 7.87 then yields $f(x) := \Gamma(x)$. \square

7.92 $\Gamma(1/2) = \sqrt{\pi}$. The substitution $t = \sin^2 \theta$ in the definition of the beta function turns (7.53) into

$$\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} = 2 \int_0^{\pi/2} (\sin \theta)^{2x-1} (\cos \theta)^{2y-1} d\theta.$$

This for $x = y = 1/2$ gives

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}. \quad (7.54)$$

7.93 $\int_0^\infty e^{-x^2} dx = \sqrt{\pi}$. The substitution $t = s^2$ in the definition of Γ yields

$$\Gamma(x) := 2 \int_0^\infty s^{2x-1} e^{-s^2} ds.$$

The special case $x = 1/2$ then gives the important

$$\int_{-\infty}^{+\infty} e^{-s^2} ds = 2 \int_0^\infty e^{-s^2} ds = \sqrt{\pi}.$$

7.94 Duplication formula of Legendre. It is not difficult to verify that Theorem 7.89 applies to the function

$$f(x) := \frac{2^{x-1}}{\sqrt{\pi}} \Gamma\left(\frac{x}{2}\right) \Gamma\left(\frac{x+1}{2}\right);$$

this yields the so-called *duplication formula of Legendre*

$$\Gamma(x) \Gamma\left(x + \frac{1}{2}\right) = 2^{1-2x} \sqrt{\pi} \Gamma(2x).$$

7.95 Formula of complementary arguments. Performing the change of variable $t = s/(1+s)$, we get

$$B(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt = \int_0^{+\infty} \frac{s^{x-1}}{(1+s)^{x+y}} ds.$$

In particular if $0 < x < 1$,

$$\Gamma(x) \Gamma(1-x) = B(x, 1-x) = \int_0^{\infty} \frac{s^{x-1}}{1+s} ds.$$

If $x = \frac{2m+1}{2n}$, $m > n$, performing another change of variable, $t = s^{1/2n}$,

$$\int_0^{\infty} \frac{s^{\frac{2m+1}{2n}-1}}{1+s} ds = 2n \int_0^{\infty} \frac{x^{2m}}{1+x^{2n}} ds = \pi \frac{1}{\sin\left(\frac{2m+1}{2n}\pi\right)};$$

see Example 5.36. Therefore,

$$\Gamma(x) \Gamma(1-x) = B(x, 1-x) = \frac{\pi}{\sin(\pi x)}$$

if $x = (2m+1)/(2n)$, $n > m$. Since $\{(2m+1)/2n \mid n > m\}$ is dense in $]0, 1[$, and Γ is continuous, we conclude

$$\Gamma(x) \Gamma(1-x) = B(x, 1-x) = \frac{\pi}{\sin(\pi x)}$$

also for any $x \in]0, 1[$.

7.96. Taking logarithms on both sides of (7.51) we get

$$\log \Gamma(x) = -\log x - \gamma x - \sum_{n=1}^{\infty} \left(\log \left(1 + \frac{x}{n} \right) - \frac{x}{n} \right). \quad (7.55)$$

Expanding the logarithms occurring in the infinite series we get

$$\log \Gamma(x) = -\log x - \gamma x + \sum_{n=1}^{\infty} \sum_{j=2}^{\infty} (-1)^j \frac{1}{j} \left(\frac{x}{n} \right)^j$$

and hence, by Weierstrass's double series theorem and the relation $\Gamma(x+1) = x\Gamma(x)$,

$$\log \Gamma(x+1) = -\gamma x + \sum_{m=2}^{\infty} \frac{(-1)^m}{m} \zeta(m) x^m$$

with $\zeta(m) := \sum_{n=1}^{\infty} \frac{1}{n^m}$.

7.97 The function ψ . The *Gaussian psi function* or *digamma* is defined as the logarithmic derivative of the gamma function

$$\psi(x) = D \log \Gamma(x) = \frac{\Gamma'(x)}{\Gamma(x)}.$$

From (7.55) we obtain

$$\psi(x) + \gamma = -\frac{1}{x} - \sum_{k=1}^{\infty} \left(\frac{1}{x+k} - \frac{1}{k} \right),$$

the series being absolutely and uniformly convergent in any bounded closed interval of $]0, \infty[$. In particular we have

$$\psi(1) = -\gamma$$

since $\sum_{k=1}^{\infty} (1/(1+k) - 1/k) = -1$. By logarithm differentiation we can easily translate relations of the Γ function into relations for the ψ . For instance, we have

$$\psi(x+1) - \psi(x) = \frac{1}{x}$$

and therefore for any $n \in \mathbb{N}$, $n \geq 1$,

$$\psi(x+n) - \psi(x) = \sum_{k=1}^n \frac{1}{x+k-1};$$

also

$$\psi(x) - \psi(1-x) = -\pi \cot \pi x,$$

$$2 \log 2 + \psi(x) + \psi\left(x + \frac{1}{2}\right) = 2\psi(2x),$$

in particular

$$\psi(1/2) = -\gamma - 2 \log 2,$$

$$\psi(n+1) = -\gamma + \sum_{k=1}^n \frac{1}{k},$$

$$\psi(n+1/2) = -\gamma - 2 \log 2 + 2 \sum_{k=1}^n \frac{1}{2k-1}.$$

7.98 An integral representation of ψ . We conclude this section by giving an integral representation of ψ . We have

$$\begin{aligned}\psi(x) + \gamma &= -\frac{1}{x} - \sum_{k=1}^{\infty} \left(\frac{1}{x+k} - \frac{1}{k} \right) \\ &= -\int_0^{\infty} e^{-tx} dx - \sum_{k=1}^{\infty} \int_0^{\infty} \left(e^{-t(k+x)} - e^{-tx} \right) dx.\end{aligned}$$

Reversing the order of summation and integration and using the formula for the sum of a geometric series we then can easily conclude

$$\psi(x) + \gamma = \int_0^{\infty} \frac{e^{-t} - e^{-tx}}{1 - e^{-t}} dt.$$

7.99 An integral representation of ψ' . Finally, the formula for ψ' is very useful. It is obtained by differentiating under the integral sign

$$\psi'(x) = \int_0^{\infty} \frac{t}{1 - e^{-t}} dt. \quad (7.56)$$

Actually, in the above, reversing the order of summation and integration and differentiating under the integral sign require some justification. One can provide ad hoc justification, but we prefer not doing it since it becomes much simpler in the context of *Lebesgue integration*.

c. Asymptotics of Γ and ψ

Suppose that $f(t) : [0, \infty[\rightarrow \mathbb{R}$ is a smooth function which has, together with its derivatives, at most a polynomial growth near infinity. Then

$$\varphi(x) := \int_0^{\infty} e^{-xt} f(t) dt, \quad x > 0$$

is well defined. We are interested in its asymptotic expansion near infinity. Integrating by parts n times we infer

$$\begin{aligned}\varphi(x) &= -\sum_{k=0}^n \frac{D^k f(t) e^{-xt}}{x^{k+1}} \Big|_0^{\infty} + \frac{1}{x^{n+1}} \int_0^{\infty} e^{-xt} D^{n+1} f(t) dt \\ &= \sum_{k=0}^{\infty} \frac{D^k f(0)}{x^{k+1}} + \frac{r_n(x)}{x^{n+1}}.\end{aligned}$$

If we also assume that

$$\int_0^{\infty} |D^{n+1} f(t)| dt < +\infty,$$

we readily conclude

$$\varphi(x) = \sum_{k=0}^{n-1} \frac{D^k f(0)}{x^{k+1}} + O\left(\frac{1}{x^{n+1}}\right) \quad \text{as } x \rightarrow \infty.$$

The previous remarks apply, as it is easily verified, to the derivative of the ψ function

$$\psi'(x) = \int_0^\infty e^{-xt} \left(t + \frac{t}{e^t - 1} \right) dt;$$

as $D^k(t/(e^t - 1)) = B_k$, we then conclude

$$\psi'(x) = \frac{1}{x} + \frac{1}{2x^2} + \sum_{k=1}^{n-1} \frac{B_{2k}}{x^{2k+1}} + O\left(\frac{1}{x^{2n+1}}\right).$$

Integrating twice over $]0, \infty[$, we obtain

$$\psi(x) = A + \log x - \frac{1}{2x} - \sum_{k=1}^{n-1} \frac{B_{2k}}{2k x^{2k}} + O\left(\frac{1}{x^{2n}}\right), \quad (7.57)$$

$$\begin{aligned} \log \Gamma(x) &= B + (A - 1)x + \left(x - \frac{1}{2}\right) \log x \\ &+ \sum_{k=1}^{n-1} \frac{B_{2k}}{2k(2k-1)x^{2k-1}} + O\left(\frac{1}{x^{2n-1}}\right) \end{aligned} \quad (7.58)$$

where A and B are two constants. In particular we have

$$\log \Gamma(x) = B + (A - 1)x + (x - 1/2) \log x + O\left(\frac{1}{x}\right) \quad \text{as } x \rightarrow \infty.$$

From the relation $\Gamma(x+1) - \Gamma(x) - \log x = 0$ it follows that

$$A - 1 + \left(x + \frac{1}{2}\right) \log(x+1) - \left(x - \frac{1}{2}\right) \log x = O\left(\frac{1}{x}\right) \quad \text{as } x \rightarrow \infty,$$

which implies $A = 0$. Similarly from the duplication formula of Legendre we infer that $B = (\log 2\pi)/2$. Therefore we conclude with the *asymptotic representations* for Γ and ψ : For all $n \in \mathbb{N}$, $n \geq 1$, we have, when $n \rightarrow \infty$,

$$\psi(x) = \log x - \frac{1}{2x} - \sum_{k=1}^{n-1} \frac{B_{2k}}{2k x^{2k}} + O\left(\frac{1}{x^{2n}}\right), \quad (7.59)$$

$$\begin{aligned} \log \Gamma(x) &= x \log x - x - \log \sqrt{x} + \log \sqrt{2\pi} \\ &+ \sum_{k=1}^{n-1} \frac{B_{2k}}{2k(2k-1)x^{2k-1}} + O\left(\frac{1}{x^{2n-1}}\right); \end{aligned} \quad (7.60)$$

in particular

$$\Gamma(x+1) = \sqrt{2\pi x} \left(\frac{x}{e}\right)^x \left(1 + \frac{1}{12x} + O\left(\frac{1}{x^2}\right)\right)$$

and, for $x = n$, *Stirling's formula*,

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n.$$

7.5 Summing Up

Convergence and continuity of the sum

- The radius of convergence ρ of a power series $\sum_{n=0}^{\infty} a_n z^n$ is defined by

$$\frac{1}{\rho} := \limsup_{n \rightarrow \infty} \sqrt[n]{|a_n|},$$

where we have adopted the conventions $1/\infty = 0$, $1/0^+ = \infty$. If $\rho > 0$, equivalently, if the sequence $\{|a_n|\}$ grows at most exponentially, then $\sum_{n=0}^{\infty} a_n z^n$ converges absolutely in the interior of the *disc of convergence* $\{z \in \mathbb{C} \mid |z| < \rho\}$ and does not converge if $|z| > \rho$. Therefore the domain $\Delta \subset \mathbb{C}$ in which $\sum_{n=0}^{\infty} a_n z^n$ converges, that is the domain in which the sum $S(z) = \sum_{n=0}^{\infty} a_n z^n$ exists as a complex number, is the disc of convergence union eventually part of or the whole boundary $\{z \in \mathbb{C} \mid |z| = \rho\}$.

- Powers series *converge uniformly* on any disc $\{z \in \mathbb{C} \mid |z| \leq r\}$, $\forall r < \rho$. This means that for any $r < \rho$ the error we get substituting the sum with a partial sum

$$\left| \sum_{n=p}^{\infty} a_n z^n \right|$$

is bounded by a quantity $c(k, r)$ that goes to zero as $p \rightarrow \infty$ and is independent of z provided $|z| \leq r$. This is equivalent to saying that

$$M(p, r) := \sup_{z \in [-r, r]} \left| \sum_{n=p}^{\infty} a_n z^n \right| \rightarrow 0 \quad \text{as } p \rightarrow \infty.$$

Uniform convergence on all discs strictly included in $|z| < \rho$, is less than the uniform convergence on $\{z \mid |z| < \rho\}$. However, it suffices to prove that

- the sum $S(z) := \sum_{n=0}^{\infty} a_n z^n$ is continuous on the interior of the disc of convergence.

Differentiation and integration of power series

Let $\sum_{n=0}^{\infty} a_n z^n$ be a complex power series with a positive radius of convergence $\rho > 0$, and sum $S(z)$. Then

- $S(z)$ has a complex derivative on $\{z \in \mathbb{C} \mid |z| < \rho\}$ and $DS(z) = \sum_{n=1}^{\infty} n a_n z^{n-1}$, that is, *the sum of power series can be differentiated term by term in the interior of the disc of convergence*.
- Actually $S(z)$ has complex derivatives of any order on $\{z \mid |z| < \rho\}$ and

$$D^k S(z) = \sum_{n=1}^{\infty} n(n-1) \cdots (n-k+1) z^{n-k}.$$

- We have $a_k := D^k S(0)/k!$ $\forall k$, that is, *each power series with a positive radius of convergence is the Taylor series of its sum*.
- *The sum of a power series can be integrated term by term in the interior of the disc of convergence*: if $\rho > 0$ denotes the radius of the disc of convergence of $\sum_{n=0}^{\infty} a_n z^n$, then, if $|z| < \rho$, we have

$$\int_0^z S(z) dz = \sum_{n=0}^{\infty} a_n \frac{z^{n+1}}{n+1}.$$

The symbol \int_0^z is the classical oriented definite integral in case $z \in \mathbb{R}$, while it is suitably defined when $z \in \mathbb{C}$, see Section 7.1.3.

Boundary values

Let $\sum_{n=0}^{\infty} a_n z^n$ be a complex power series with a positive radius of convergence $\rho > 0$, and sum $S(z)$. Two cases can occur: either the series converges absolutely at a point z_0 with $|z_0| = \rho$, or the series eventually converges, but not absolutely, at some points with $|z| = \rho$. In the first case, which implies the uniform convergence of the series and the continuity of the sum on the closed disc $\{z \mid |z| \leq \rho\}$, the convergence test of Proposition 7.29 may be useful. In the second case, some information at the boundary is provided by Dirichlet's and Abel's theorems, Theorems 7.30 and 7.31. A consequence of Abel's theorem is that $\Delta_1 \subset \Delta_2 \subset \Delta_3$ if Δ_1, Δ_2 and Δ_3 are respectively the domains of convergence of $\sum_{n=1}^{\infty} n a_n z^{n-1}$, $\sum_{n=0}^{\infty} a_n z^n$ and $\sum_{n=0}^{\infty} a_n \frac{z^{n+1}}{n+1}$.

7.6 Exercises

7.100 ¶. Show that

$$\begin{aligned} \sum_{n=0}^{\infty} (n+1)z^n &= \frac{1}{(1-z)^2}, & \sum_{n=0}^{\infty} n z^n &= \frac{z}{1-z^2}, \\ \sum_{n=0}^{\infty} n^2 z^n &= \frac{z^2 + z}{(1-z)^3}, & \sum_{n=0}^{\infty} (-1)^n z^n &= \frac{1}{1+z}, \\ \sum_{n=0}^{\infty} z^{2n} &= \frac{1}{1-z^2}, & \sum_{n=0}^{\infty} \frac{(-1)^n}{n!} &= \frac{1}{e}, \\ \sum_{n=0}^{\infty} \frac{(-1)^n}{2^n} &= \frac{2}{3}. \end{aligned}$$

7.101 ¶. Compute the sums of the following series

$$\begin{aligned} \sum_{n=0}^{\infty} \frac{\cos(1+2n)}{n!}, & \quad \sum_{n=0}^{\infty} (-1)^n \frac{\sin(1+2n)}{(2n)!}, \\ \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{n!}, & \quad \sum_{n=0}^{\infty} \frac{z^{3n+2}}{n!}, \\ \sum_{n=0}^{\infty} \frac{z^{2n}}{2n+1}, & \quad \sum_{n=0}^{\infty} \frac{z^{2n+1}}{n}, \\ \sum_{n=0}^{\infty} \frac{z^{2n+1}}{n+1}, & \quad \sum_{n=0}^{\infty} \frac{z^n}{n(n+1)}, \\ \sum_{n=0}^{\infty} \frac{z^n}{n^2}, & \quad \sum_{n=1}^{\infty} \log(\cos(x/2^k)). \end{aligned}$$

[Hint: For the last series, show that

$$\prod_{k=1}^n \cos\left(\frac{x}{2^k}\right) = \frac{\sin x}{2^n \sin(x/2^n)}.]$$

$$\begin{aligned}
e^z &= \sum_{n=0}^{\infty} \frac{z^n}{n!}, & z \in \mathbb{C}, \\
\log(1+z) &= \sum_{n=0}^{\infty} (-1)^n \frac{z^{n+1}}{n+1}, & |z| \leq 1, z \neq -1, \\
\sin z &= \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)!}, & z \in \mathbb{C}, \\
\sinh z &= \sum_{n=0}^{\infty} \frac{z^{2n+1}}{(2n+1)!}, & z \in \mathbb{C}, \\
\cos z &= \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!}, & z \in \mathbb{C}, \\
\cosh z &= \sum_{n=0}^{\infty} \frac{z^{2n}}{(2n)!}, & z \in \mathbb{C}, \\
\arcsin z &= \sum_{n=0}^{\infty} \frac{(2n-1)!!}{(2n)!!} \frac{z^{2n+1}}{2n+1}, & |z| \leq 1, \\
\sinh^{-1} z &= \sum_{n=0}^{\infty} (-1)^n \frac{(2n-1)!!}{(2n)!!} \frac{z^{2n+1}}{2n+1}, & |z| \leq 1, z \neq \pm 1, \\
\arccos z &= \frac{\pi}{2} - \arcsin z, & |z| \leq 1, \\
\arctan z &= \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{2n+1}, & |z| \leq 1, z \neq \pm i, \\
\tanh^{-1} z &= \sum_{n=0}^{\infty} \frac{z^{2n+1}}{2n+1}, & |z| \leq 1, z \neq \pm 1, \\
\frac{1}{1-z} &= \sum_{n=0}^{\infty} z^n, & |z| < 1, \\
(1+z)^\alpha &= \sum_{n=0}^{\infty} \binom{\alpha}{n} z^n, & |z| < 1,
\end{aligned}$$

where

$$(2n)!! = \begin{cases} 1 & \text{if } n = 0, \\ 2n(2n-2) \cdots 4 \cdot 2 & \text{if } n \geq 1, \end{cases} \quad (2n+1)!! = (2n+1)(2n-1) \cdots 5 \cdot 3 \cdot 1$$

and for $\alpha \in \mathbb{R}$

$$\binom{\alpha}{n} := \frac{\alpha(\alpha-1)(\alpha-2) \cdots (\alpha-n+1)}{n!}.$$

Figure 7.4. A table of Taylor series of some elementary functions.

7.102 ¶. Let $f(x) := |\sin x|/x$, $x > 0$, and for $n = 0, 1, \dots$, let

$$f_n(x) := \begin{cases} f(x) & \text{if } n\pi \leq x < (n+1)\pi, \\ 0 & \text{otherwise.} \end{cases}$$

Show that

- (i) $\sum_{n=0}^{\infty} f_n(x) = f(x) \quad \forall x \geq 0$,
- (ii) $\sup_{x \geq 0} \left| \sum_{k=n+1}^{\infty} f_k(x) \right| \rightarrow 0$,
- (iii) $\sum_{n=0}^{\infty} \sup_{x \geq 0} |f_n(x)| = +\infty$.

7.103 ¶. Show that $\sum_{n=1}^{\infty} (x^n + (-1)^{n+1}/n)$ converges uniformly in $[0, 1/2]$, but it does not converge absolutely.

7.104 ¶. Show that $x + \sum_{k=0}^{\infty} (kxe^{-kx^2} - (k+1)xe^{-(k+1)x^2})$ converges, the limits of the sum are the sum of the limits, but it is not uniformly convergent.

7.105 ¶. Show the following

Proposition. Suppose $\sum_{n=0}^{\infty} a_n z^n$ converges uniformly on $|z| = 1$. Then $\sum_{n=0}^{\infty} a_n z^n$ converges uniformly on $|z| \leq 1$.

[Hint: Reread the proof of Abel's theorem.]

7.106 ¶¶. Let $f(z) = \sum_{n=0}^{\infty} a_n z^n$ on $|z| < r$. Then $f(z)$ is representable as power series $\sum_{n=0}^{\infty} b_n (z - z_0)^n$ with center any z_0 with $|z_0| < r$ and domain of convergence that contains $\{z \mid |z - z_0| < r\}$. [Hint: Set $z = z_0 + h$ and $\rho := |z_0| - |h|$, and write

$$\sum_{k=0}^{\infty} a_k \rho^k = \sum_{k=0}^{\infty} |a_k| \left(\sum_{t=0}^k \binom{k}{t} |z_0|^{k-t} |h|^t \right).$$

7.107 ¶¶ Composition of power series. Suppose that $S(x) = \sum_{n=0}^{\infty} a_n x^n$ and $T(y) := \sum_{n=0}^{\infty} b_n y^n$ are two power series with $T(0) = 0$ and, respectively, with positive radii of convergence $\rho(S)$ and $\rho(T)$. Show that the composition $S \circ T$ is the sum of a power series with positive radius of convergence. More precisely show that, if $r > 0$ is such that $\sum_{n=0}^{\infty} |b_n| r^n < \rho(S)$, then the radius of convergence of $S \circ T$ is at least r .

7.108 ¶¶ Inverse of a power series. Let $S(x) = \sum_{n=1}^{\infty} a_n x^n$ be a power series with $S(0) \neq 0$ and positive radius of convergence. Show that there is a power series T with radius of convergence 1 such that $S(x)T(x) = 1$.

7.109 ¶¶ Reciprocal power series. Let $S(x) = \sum_{n=0}^{\infty} a_n x^n$ be a power series with $S(0) = 0$, $S'(0) \neq 0$ and positive radius of convergence. Show that there is a power series T with $T(0) = 0$ and positive radius of convergence such that $S(T(x)) = x$ [Hint: see e.g., Cartan, *Théorie élémentaire des fonctions analytiques d'une ou plusieurs variables complexes*, Hermann, Paris, 1961.]

7.110 ¶¶. Let $f: \mathbb{N} \rightarrow \mathbb{R}$ be a function such that $f(1) \neq 0$, $f(x_1 x_2) = f(x_1)f(x_2)$ $\forall x_1, x_2$ and $\sum_{n=1}^{\infty} |f(n)| < \infty$. Show that

- (i) $f(1) + \sum_{k=1}^{\infty} |f(n)|^k < \infty$, $\prod_{p=1}^{\infty} (1 - f(p)) < \infty$, $\prod_{p=1}^{\infty} \frac{1}{1 - f(p)} < \infty$ and finally

$$\frac{1}{1 - f(p)} = \sum_{k=0}^{\infty} f^k(p) = 1 + f(p) + f^2(p) + \dots$$

(ii) If $\{p_n\}$ is the sequence of primes and

$$P_n := \frac{1}{1-f(p_1)} \frac{1}{1-f(p_2)} \cdots \frac{1}{1-f(p_n)} = \prod_{r=1}^n \left(1 + \sum_{k=1}^{\infty} f(p_r)^k\right),$$

then

$$P_n = \sum_N f(N),$$

the sum being taken on all naturals N which decompose in prime factors containing only the primes p_1, p_2, \dots, p_n .

(iii) *Euler's formula*

$$\sum_{n=1}^{\infty} f(n) = \prod_{p \text{ prime}} \frac{1}{1-f(p)}.$$

(iv) If $f(n) = 1/n^s$, with $s > 1$, then

$$\sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_{p \text{ prime}} \frac{1}{1-p^{-s}},$$

the product being taken on all primes. The function

$$\zeta(s) := \sum_{n=1}^{\infty} \frac{1}{n^s}$$

is actually well defined for all $s \in \mathbb{C}$ with $\Re s > 1$ and is called *Riemann's ζ -function*.

7.111 ¶¶ Euler. Let $\{p_i\}$ be the sequence of primes. Then $\sum_{i=1}^{\infty} \frac{1}{p_i} = +\infty$. [Hint: If $\sum_{i=1}^{\infty} 1/p_i < \infty$, then for some $m \in \mathbb{N}$ we have $\sum_{i>m} 1/p_i < 1/2$. Setting $\alpha := \prod_{i=1}^m p_i$, p_i divides $1 + n\alpha$ at most for $i > m$, hence

$$\sum_{n=1}^{\infty} \frac{1}{1+n\alpha} \leq \sum_{\ell=1}^{\infty} \left(\sum_{i>m} \frac{1}{p_i} \right)^{\ell} \leq \sum_{\ell=1}^{\infty} \frac{1}{2^{\ell}} = 1.]$$

7.112 ¶. Find the series solutions of the differential equations

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

7.113 ¶. Find the series solutions of the Cauchy problems

$$\begin{cases} y'' + (x-1)y' - (x-1)y = 0, \\ y(1) = 1, \quad y'(1) = 0, \end{cases} \quad \begin{cases} (1+x^2)y'' + y = 0, \\ y(0) = y'(0) = 1. \end{cases}$$

7.114 ¶. The equation

$$y'' + e^x y = 0$$

has a solution of the form $y(x) = \sum_{k=0}^{\infty} a_k x^k$ that satisfies $y(0) = 1$, $y'(0) = 0$. Find some of its first terms.

7.115 ¶ Legendre's equation. Find the series solutions of the ODE

$$(1-x^2)y'' - 2xy' + \lambda y = 0$$

called *Legendre's polynomials*. [Hint: $y(x) = a_0 \left(1 - \frac{\lambda}{2}x^2 - \frac{6\lambda}{4 \cdot 3 \cdot 2}x^4 + \cdots\right) + a_1 \left(x + \frac{2\lambda}{3 \cdot 2}x^3 + \frac{(12-\lambda)(2-\lambda)}{5 \cdot 4 \cdot 3 \cdot 2}x^5 + \cdots\right).$]

7.116 ¶ Hermite polynomials. Find the series solutions of the ODE

$$y'' - 2xy' + 2y = 0,$$

called *Hermite's polynomials*.

7.117 ¶ Euler equation. Show that there is no series solution except 0 of the equation

$$x^2 y'' + \alpha x y' + \beta y = 0.$$

Show that $y(x) = x^\gamma$ where γ satisfies $\gamma(\gamma - 1) + \alpha\gamma + \beta = 0$, is a solution.

7.118 ¶ Frobenius's method. Examples 7.53 and 7.54 suggest that the power series method may not work if the higher order coefficient of the second order linear ODE vanishes at $x = 0$. In this case we may try solutions of the form

$$x^\gamma \sum_{k=0}^{\infty} a_k x^k, \quad \gamma \in \mathbb{R}.$$

Try the method with the following *Bessel's equation*

$$x^2 y'' + x y' + (x^2 - \frac{1}{4}) y = 0$$

and *Laguerre's equation*

$$x y'' + (1 - x)y' + y = 0.$$

7.119 ¶. Let $A(x)$ and $E(x)$ be respectively the enumerator and the exponential enumerator of $\{a_n\}$. Show that

$$A(x) = \int_0^\infty e^{-sx} E(sx) ds.$$

7.120 ¶. Let $\{p_n\}$ be a sequence in $[0, 1[$ and let $P(x)$ be the enumerator of $\{p_n\}$. The k -moment of $\{p_n\}$ is defined by

$$m_k := \sum_{j=0}^{\infty} j^k p_j.$$

Assuming that m_k is finite for all $k \geq 0$, show that the enumerator of $\{m_k\}$ is

$$M(x) = P(e^x).$$

7.121 ¶. Show that the binary numbers of $2n$ bits are $\binom{2n}{n}$.

7.122 ¶. Show that $\sum_{r=0}^n r \binom{n}{r} = n2^{n-1}$.

7.123 ¶. Show that

$$\frac{1}{(1-z)^{m+1}} = \sum_{k=0}^{\infty} \binom{m+k}{m} z^k, \quad \frac{z^m}{(1-z)^m} = \sum_{k=0}^m \binom{k}{m} z^k.$$

7.124 ¶. Show that

$$e^{2tx-t^2} = \sum_{n=0}^{\infty} \frac{H_n(x)}{n!} t^n$$

where $H_n(x)$ solves $y'' - 2xy' + 2ny = 0$.

7.125 ¶. Show that the enumerator for the selection of r objects out of n objects, $r \geq n$, with unlimited repetitions but with each object included in the selection, is

$$\left(\sum_{k=1}^{\infty} x^k\right)^n = \sum_{r=n}^{\infty} \binom{r-1}{n-1} x^r.$$

7.126 ¶¶. Show that the number of ways in which r nondistinct objects can be distributed in n distinct cells, with the condition that no cell contains less than q nor more than $q+z-1$ objects, is the coefficient of x^{r-qn} in the expansion of $\left((1-x^z)/(1-x)\right)^n$.

7.127 ¶¶. Show that a convex polygon of $n+2$ sides can be divided into

$$c_n := \frac{1}{n+1} \binom{2n}{n}$$

triangles by means of diagonals that do not intersect. The numbers c_n are called *Catalan numbers*. [Hint: Notice that $c_{n+1} = \sum_{k=0}^{\infty} c_k c_{n-k}$, hence, if $c(x) = \sum_{n=0}^{\infty} c_n x^n$, we have $c^2(x) = \sum_{n=0}^{\infty} c_{n+1} x^n$ and $xc^2(x) = c(x) - 1$.]

7.128 ¶¶. Show that the exponential enumerator for the distribution of r or less objects into n distinct cells, with objects in the same cell ordered, is $\exp x/(1-x)$.

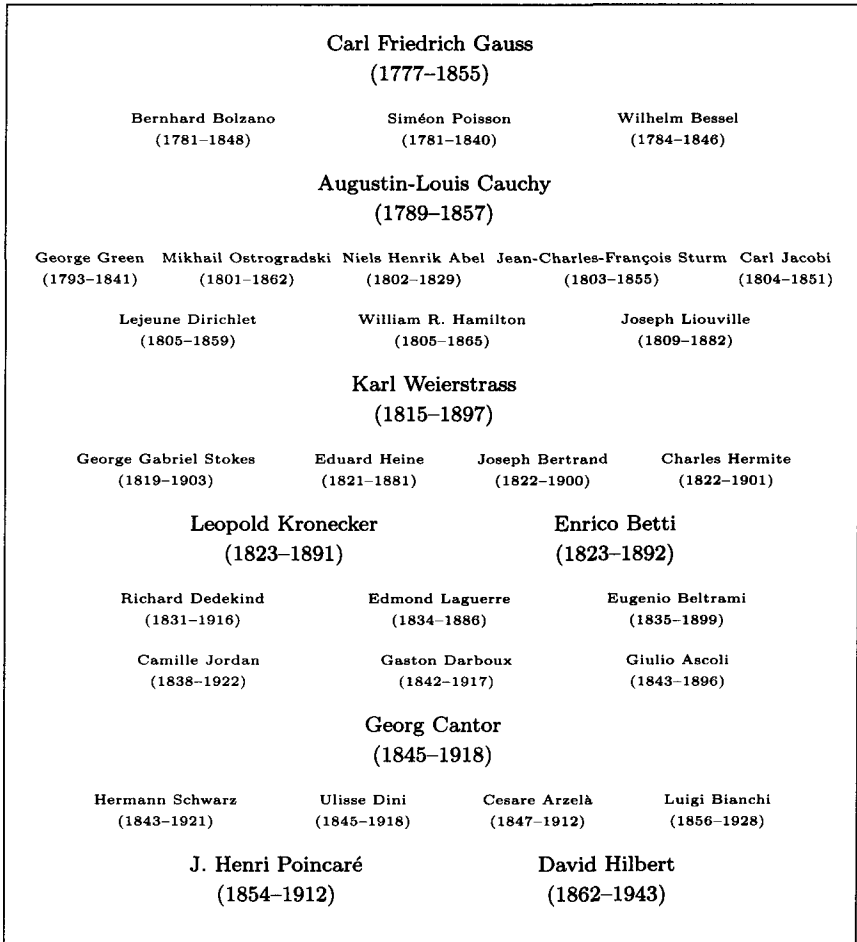


Figure 7.5. Infinitesimal analysis: a chronology from Gauss to Poincaré and Hilbert.

8. Discrete Processes

The laws of classical physics are *deterministic*: if we know *exactly* the state of a system at a given instant, we know its state for all times. Such a principle, which mathematically corresponds to the *existence and uniqueness theorem for the Cauchy problem*, has been (and is) a key idea in scientific thought. Pierre-Simon Laplace (1749–1827) wrote in his *Essai philosophique sur les probabilités*.

Nous devons donc envisager l'état présent de l'Univers comme l'effet de son état antérieur, et comme cause de celui qui va suivre. Une intelligence qui pour un instant donné connaîtrait toutes les forces dont la nature est animée et la situation respective des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre ses données à l'analyse, embrasserait dans la même formule les mouvements des plus grands corps de l'Univers et ceux du plus léger atome : rien ne serait incertain pour elle, et l'avenir, comme le passé, serait présent à ses yeux. L'esprit humain offre dans la perfection qu'il a su donner à l'astronomie une faible esquisse de cette intelligence.¹

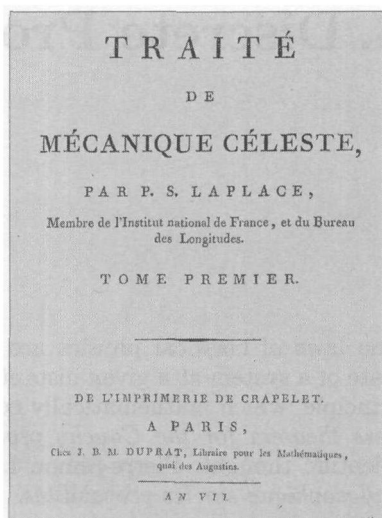
But this principle is often contradicted by everyday experience, when some facts seem to take place unpredictably and at random, as is the case with meteorology.

From the point of view of *predictability*, since there will always be a certain degree of uncertainty on the initial situation, things will be predictable if two initially close states evolve closely, otherwise one has to expect chaotic behavior: close states evolve into paths that are very far from each other. Probably the first to have stated precisely the phenomenon of *sensitive dependence on initial conditions* were Jacques Hadamard (1865–1963), who studied the flux of geodesic lines on a surface, and Pierre Duhem

¹ Therefore we have to consider the present state of the universe as the effect of its previous state and cause of its future state. An intelligence that at a given instant could know all the forces that animate nature and the respective situations of all the beings that constitute it, an intelligence that, moreover, could be large enough to be able to analyze all these data, could contain in the same formula both the movements of the largest bodies in the universe and of the smallest atom: nothing would be uncertain for it and the future, as well as the past, would be present to its eyes. The human spirit offers just a feeble trace of this intelligence in the perfection it was able to give to astronomy.



Figure 8.1. Pierre-Simon Laplace (1749–1827) and the frontispiece of his *Mécanique céleste*.



(1861–1916), who observed how the sensitive dependence on initial conditions made long term previsions illusory for the systems considered by Hadamard². The fact that these systems are no exception seems clear to J. Henri Poincaré (1854–1912) who writes in *Science et méthode*

Une cause très petite, qui nous échappe, détermine un effet considérable que nous ne pouvons pas ne pas voir, et alors nous disons que cet effet est dû au hasard. Si nous connaissions exactement les lois de la nature et la situation de l'Univers à l'instant initial, nous pourrions prédire exactement la situation de ce même Univers a un instant ulterieur. Mais, lors même que les lois naturelles n'auraient plus de secret pour nous, nous ne pourrions connaître la situation initiale qu'approximativement. Si cela nous permet de prévoir la situation ulterieure avec la même approximation, c'est tout ce qu'il nous faut, nous disons que le phénomène a été prévu, qu'il est régi par des lois; mais il n'en est pas toujours ainsi, il peut arriver que de petites différences dans les conditions initiales en engendrent de très grandes dans les phénomènes finaux; une petite erreur sur les premières produirait une erreur énorme sur les derniers. La prédiction devient impossible et nous avons le phénomène fortuit.³

² See *La théorie physique, son objet et sa structure*, Éditions Chevalier et Rivière, 1906.

³ A very small cause that escapes our attention determines a notable effect that we cannot fail to see, and in this case we say that it is due to hazard. If we knew exactly the laws of nature and the situation of the universe at the initial moment, we could

He also adds

Comment devons-nous nous représenter un récipient rempli de gaz ? D'innombrables molécules, animées de grande vitesses, sillonnent ce récipient dans tous les sens ; à chaque instant elles choquent les parois, ou bien elles se choquent entre elles ; et ces chocs ont lieu dans les conditions les plus diverses. Ce qui nous frappe surtout ici, ce n'est pas la petitesse des causes, c'est leur complexité. Et cependant, le premier élément se retrouve encore ici et joue un rôle important. Si une molécule était déviée vers la gauche ou vers la droite de sa trajectoire, d'une quantité très petite, comparable au rayon d'action des molécules gazeuses, elle éviterait un choc, ou elle le subirait dans des conditions différentes, et cela ferait varier, peut-être de 90° ou de 180° , la direction de sa vitesse après le choc.

Et ce n'est pas tout, il suffit, nous venons de le voir, de dévier la molécule avant le choc d'une quantité infiniment petite, pour qu'elle soit déviée, après le choc, d'une quantité finie.⁴

This somehow explains how a deterministic and regular behavior may generate chaos; but on the other hand it may suggest that a chaotic behavior may create order, possibly on a different scale from the macroscopic behavior of the gas.

Chaotic behavior may be generated by sensitive dependence on the parameters of the system, as well as by sensitive dependence on the initial data. For instance, the behavior of water coming out of a slightly open tap is regular, while it gets chaotic if the tap is completely open. In the same way, the behavior of a fluid between two rotating cylinders is regular when they rotate and gets more and more chaotic as the rotation speed increases.

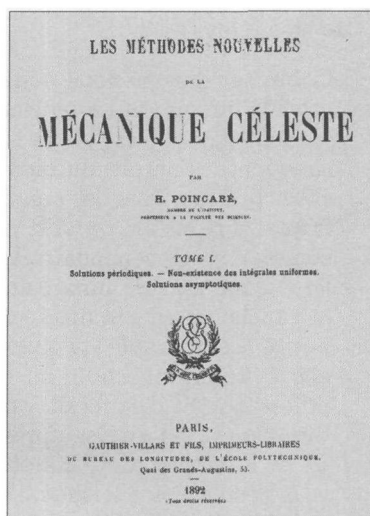
Poincaré wrote also

predict exactly the situation of the same universe at a succeeding moment, but even if it were the case that the natural laws had no longer any secret for us, we could still only know the initial situation approximately. If that enables us to predict the succeeding situation with *the same approximation*, that is all we require, and we should say that the phenomenon has been predicted, that it is governed by laws. But it is not always so: it may happen that *small differences in the initial conditions produce very great ones in the final phenomena*. A small error in the former will produce an enormous error in the latter. Prediction becomes impossible, and we have the fortuitous phenomenon.

⁴ How should we represent a container full of gas? Innumerable molecules race inside the container in all directions; at every instant they hit the container's sides or collide with one another; and all these collisions take place in the utmost diverse conditions. What strikes us in this case is not the smallness of the causes but mainly their complexity. And yet, the first element is still present and plays an important role. If a molecule were to be diverted to its left or right by a small quantity comparable to the range of action of a gas molecule, it could avoid a collision or undergo the collision under different conditions, and this could change its direction by 90° or maybe 180° degrees. And this is not all, we have just seen that it is sufficient to divert the molecule of an infinitesimal quantity before the collision in order to divert it, after the collision, of a finite quantity.



Figure 8.2. J. Henri Poincaré (1854–1912) and the frontispiece of his *Méthodes nouvelles de mécanique céleste*.



Pourquoi les météorologistes ont-ils tant de peine à prédire le temps avec quelque certitude? Pourquoi les chutes de pluie, les tempêtes elles-mêmes nous semblent-elles arriver au hasard, de sorte que bien de gens trouvent tout naturel de prier pour avoir de la pluie ou le beau temps, alors qu'il jugeraient ridicule de demander une éclipse par une prière? Nous voyons que les grandes perturbations se produisent généralement dans les régions où l'atmosphère est en équilibre instable, qu'un cyclone va naître quelque part, mais où? Ils sont hors d'état de le dire; un dixième degré en plus ou en moins en un point quelconque, le cyclone éclate ici et non pas là, et il étend ses ravages sur des contrées qu'il aurait épargnées. Si on avait connu ce dixième de degré, on aurait pu le savoir d'avance, mais les observations n'étaient ni assez serrées, ni assez précises, et c'est pour cela que tout semble dû à l'intervention du hasard.⁵

⁵ Why do meteorologists have such a hard time in foreseeing the weather with a reasonable degree of precision? Why do showers and storms seem to occur at random, so that many people find it absolutely natural to pray for rain or good weather, while they would find praying for an eclipse utterly ridiculous? We see that great perturbations generally occur in regions where the atmosphere is unstable. Meteorologists are well aware of the instability of the equilibrium and that somewhere there will be a hurricane, but where? They cannot tell, because a tenth of a degree more or less at any point will determine a hurricane here instead of there, and there will be devastations in areas that would have been spared. If one had known this tenth of a degree one could have foreseen the event, but observations were neither sufficiently frequent nor sufficiently precise, and for this reason everything seems to be due to the intervention of hazard.

Even though mathematicians have always known that dynamical systems may behave in unexpected and complicated ways, it is only with the invention of computers and the increasing interests in mathematical models for population dynamics, biology, electronic circuits with nonlinear components, astronomy and meteorology that the study of deterministic chaos has acquired importance to the point of becoming a fashionable subject not only among mathematicians and physicists. Particularly relevant in this process are the contribution of Hendrik Lorentz (1853–1928), meteorologist at MIT, who published in 1963 a simplified model of fluid that shows the rapid growth of errors in dependence on initial conditions, and the work of the two mathematicians, D. Ruelle and F. Takens, who, in 1971, conjectured that hydrodynamic turbulence may be represented by *strange attractors*, mathematical objects that describe evolutions with sensitive dependence on initial conditions.

One may have the impression that complex dynamics is typical of *dispersive* or *nonconservative* dynamics. But this is not true, as is shown by the question of the stability of the solar system.

It is commonly agreed that in Newton's opinion gravitational interactions among planets were so strong that they could compromise the stability of the system and that probably for this reason he formulated the hypothesis that it was God who controlled these instabilities in order to ensure the existence of the solar system; Newton writes in his *Principia*

It is not to be conceived that mere mechanical causes could give birth to so many regular motions. . . . This most beautiful system of the sun, planets, and comets, could only proceed from the council and dominion of an intelligent powerful Being.

During the Age of Enlightenment, Lagrange, Laplace, and Poisson provided mathematical reasons in favour of the stability of planetary orbits, showing, for instance, absence of polynomial growth in time of the major axis of the orbit up to third order with respect to the planetary masses. More recently Poincaré and George Birkhoff (1884–1944) showed that in the dynamics of planets one may encounter instabilities that make the notion in the phase space quite complex and the more recent contributions of A. N. Kolmogorov, V. A. Arnold, and J. Moser suggest the *coexistence* of both stability and instability. All the same, many questions regarding *n*-bodies are still unresolved.⁶

So far we have always referred to continuous dynamical systems. However, *discrete dynamical systems* are naturally associated to continuous ones, for instance in the form of discretization or of a Poincaré map. They also naturally occur in the study of population dynamics and are the appropriate models for computers and often show a chaotic behavior even when the corresponding continuous system does not.

⁶ See the paper by Jacques Laskar in A. Dahan Delmedico, J. L. Chabert, K. Chemla, Eds, *Chaos et Déterminisme*, Éditions du Seuil, Paris, 1992, and S. Marmi, *Chaotic behavior in the solar theory*, Séminaire Bourbaki 5^{me} année, 1998–1999, n. 854.

In this chapter we shall describe the behavior of a few simple discrete dynamical systems with the aim of showing some paths leading to chaos. In fact, a wider and more precise analysis cannot avoid a wider and more detailed study of ordinary differential equations and further technical tools. In Section 1 we discuss first and second order linear difference equations, some nonlinear examples of recurrences and continued fractions; in Section 2 we shall then illustrate some aspects of one-dimensional dynamical systems.

8.1 Recurrences

In the previous chapters we encountered on several occasions *recursive relations*, some of which lead to closed form sequences while some do not, and we studied in some detail the process of summing with the analysis of series. In this section we discuss a few more classical recurrences that from a dynamical point of view, that is from the point of view of the behavior at infinity, are quite regular.

8.1.1 Linear difference equations

In this section, we discuss first order *linear difference equations*. They are the discrete version of the first order ODE and can be solved in closed form.

a. First order linear difference equations

We recall (see Example 2.5) that, given $\{f_n\}$, the recurrence

$$\begin{cases} x_0 \text{ given,} \\ x_{n+1} = x_n + f_{n+1}, \quad \forall n \geq 0, \end{cases}$$

is equivalent to

$$x_n := x_0 + \sum_{j=1}^n f_j.$$

Moreover, we have the following.

8.1 Proposition. *Given $a \in \mathbb{R}$ and $\{f_n\}$, the solution of*

$$\begin{cases} x_0 \text{ given,} \\ x_{n+1} = ax_n + f_{n+1}, \quad \forall n \geq 0, \end{cases} \quad (8.1)$$

is

$$x_n := a^n x_0 + \sum_{j=1}^n a^{n-j} f_j. \quad (8.2)$$

In fact, the sequence $\{x_n\}$ given by (8.2) verifies the recurrence relations (8.1). The solutions of (8.1) have a structure, which is quite similar to the structure of the solutions of first order linear ODE (see, e.g., Chapter 5 of [GM1]):

- the sequence $u = \{u_n\}$, given by $u_n = a^n$ solves

$$u_0 = 1, \quad u_{n+1} = a u_n \quad n \geq 0,$$

- if $f := \{f_n\}$, then the product of convolution of u and f , $\{u * f\}$, $(u * f)_n := \sum_{j=0}^n u_{n-j} f_j$ solves

$$\begin{cases} x_0 = f_0, \\ x_{n+1} = a x_n + f_{n+1}, \quad \forall n \geq 0, \end{cases}$$

since

$$(u * f)_{n+1} = \sum_{j=0}^{n+1} u_{n+1-j} f_j = \sum_{j=0}^n a^{n+1-j} f_j + f_{n+1} = a (u * f)_n + f_{n+1};$$

- consequently, $x_n := u_n(x_0 - f_0) + (u * f)_n$ solves (8.1).

Similarly, we easily see that first order linear difference equations with varying coefficients are uniquely solvable, and we have the following.

8.2 Proposition. *Let $\{a_n\}$ and $\{f_n\}$ be two sequences. Then*

- (i) *The sequence $u := \{u_n\}$,*

$$\begin{cases} u_0 = 1, \\ u_n = \prod_{j=1}^n a_j, \quad n \geq 1, \end{cases} \quad \text{solves} \quad \begin{cases} x_0 = 1, \\ x_{n+1} = a_{n+1} x_n, \quad n \geq 0. \end{cases}$$

- (ii) *The sequence $\{x_n\}$, $x_n := u_n \sum_{j=0}^n \frac{f_j}{u_j}$, solves*

$$\begin{cases} x_0 = f_0, \\ x_{n+1} = a_{n+1} x_n + f_{n+1}, \quad n \geq 0. \end{cases}$$

- (iii) *Consequently, $\{x_n\}$,*

$$x_n := u_n(x_0 - f_0) + u_n \sum_{j=0}^n \frac{f_j}{u_j},$$

solves

$$\begin{cases} x_0 \text{ given}, \\ x_{n+1} = a_{n+1} x_n + f_{n+1}, \quad n \geq 0. \end{cases}$$

b. Second order homogeneous difference equations

A generic second order difference equation has the form

$$a x_{n+2} + b x_{n+1} + c x_n = 0, \quad (8.3)$$

where a, b and c are constants and $a \neq 0$. We are interested in finding all solutions of (8.5), or, equivalently, in solving explicitly the recurrence

$$\begin{cases} x_0 = \alpha, & x_1 = \beta, \\ a x_{n+2} + b x_{n+1} + c x_n = 0, & n \geq 0, \end{cases} \quad (8.4)$$

for any given α and $\beta \in \mathbb{R}$. As in the case of ODEs (compare, e.g., [GM1]),

- (i) if $\{x_n\}$ and $\{y_n\}$ solve (8.3), then also $\{c_1 x_n + c_2 y_n\}_n$ solves (8.3) for any $c_1, c_2 \in \mathbb{C}$,
- (ii) if λ is a solution of the *characteristic equation*

$$a\lambda^2 + b\lambda + c = 0,$$

then $\{\lambda^n\}$ solves (8.3): in fact,

$$a\lambda^{n+2} + b\lambda^{n+1} + c\lambda^n = \lambda^n(a\lambda^2 + b\lambda + c) = 0.$$

Let λ_1, λ_2 be the two solutions of the characteristic equation. Corresponding to the three cases of real and distinct roots, repeated real roots and conjugate complex roots, define the two sequences $\{u_n\}, \{v_n\}$ by

$$\begin{aligned} u_n &= \lambda_1^n, & v_n &= \lambda_2^n & \text{if } \lambda_1, \lambda_2 \in \mathbb{R}, \lambda_1 \neq \lambda_2, \\ u_n &= \lambda_1^n, & v_n &= n\lambda_1^n & \text{if } \lambda_1 = \lambda_2, \\ u_n &= |\lambda_1|^n \cos(n\varphi), & v_n &= |\lambda_1|^n \sin(n\varphi) & \text{otherwise.} \end{aligned} \quad (8.5)$$

The latter occurs if λ_1, λ_2 are complex conjugate, and in this case we have set

$$\lambda_1 = |\lambda_1| e^{i\varphi}.$$

We have the following.

8.3 Proposition. *The solution of (8.4) is the sequence $\{x_n\}$ given by $x_n = c_1 u_n + c_2 v_n$ where c_1, c_2 solves*

$$c_1 u_0 + c_2 v_0 = \alpha, \quad c_1 u_1 + c_2 v_1 = \beta. \quad (8.6)$$

Proof. (i) *Real and distinct roots.* By linearity $\{x_n\}$, $x_n = c_1 \lambda_1^n + c_2 \lambda_2^n$, solves (8.3). Moreover, since $\lambda_1 \neq \lambda_2$, for any $\alpha, \beta \in \mathbb{R}$, one can then solve for c_1, c_2 the system in (8.6).

(ii) *Complex conjugate roots.* Let $\lambda := \lambda_1$, and $\bar{\lambda} = \lambda_2$. Similarly to (i), all complex valued sequences $\{x_n\}$ $x_n = c_1 \lambda^n + c_2 \bar{\lambda}^n$, $c_1, c_2 \in \mathbb{C}$, solve (8.3). For given α, β we then solve in \mathbb{C} , since $\lambda \neq \bar{\lambda}$, (8.6) to get the solution

$$x_n = c_1 \lambda^n + c_2 \bar{\lambda}^n$$

of (8.4), an a priori complex function.

But α, β being reals, when solving (8.6), we get $c_2 = \bar{c}_1$, hence the solution found is real and written as

$$\Re(c_1 \lambda^n),$$

or, setting $c = a - ib$, $a, b \in \mathbb{R}$, by de Moivre's formula,

$$\Re(c_1 \lambda^n) = a \Re(\lambda^n) + b \Im(\lambda^n) = a |\lambda|^n \cos(n\varphi) + b |\lambda|^n \sin(n\varphi).$$

(iii) *Repeated real roots.* Let $\lambda_1 = \lambda_2 = \lambda$. Since in this case $2a\lambda + b = 0$, we have

$$\begin{aligned} a(n+2)\lambda^{n+2} - b(n+1)\lambda^{n+1} - cn\lambda^n \\ = n\lambda^n(a\lambda^2 + b\lambda + c) + \lambda^n(2a\lambda + b) = 0, \end{aligned}$$

i.e., $\{n\lambda^n\}$ solves (8.3). Therefore all sequences $\{x_n\}$, $x_n = \lambda^n(c_1 + c_2n) = c_1u_n + c_2v_n$, $c_1, c_2 \in \mathbb{R}$, solve (8.3). Since the system in (8.6) yields c_1 and c_2 , we find the solution of (8.4). \square

c. Second order nonhomogeneous difference equations

Consider the recurrence

$$\begin{cases} x_0 = \alpha, & x_1 = \beta, \\ a x_{n+2} + b x_{n+1} + c x_n = f_{n+1}, \end{cases} \quad (8.7)$$

where $a, b, c \in \mathbb{R}$ and $\{f_n\}$ is a given sequence.

8.4 Proposition. *Let $\{w_n\}$ be the sequence that solves the homogeneous recurrence*

$$\begin{cases} w_0 = 0, & w_1 = 1, \\ a w_{n+2} + b w_{n+1} + c w_n = 0, & n \geq 0. \end{cases} \quad (8.8)$$

Then the sequence $\{x_n\}$ given by

$$x_n := \frac{1}{a} \{(w * f)_n\} = \frac{1}{a} \sum_{j=0}^n w_{n-j} f_j,$$

solves

$$\begin{cases} x_0 = 0, & x_1 = f_0/a, \\ a x_{n+2} + b x_{n+1} + c x_n = f_{n+1}, & n \geq 0. \end{cases}$$

Proof. In fact,

$$\begin{aligned}
 a \sum_{j=0}^{n+2} w_{n+2-j} f_j + b \sum_{j=0}^{n+1} w_{n+1-j} f_j + c \sum_{j=0}^n w_{n-j} f_j \\
 = aw_0 f_{n+2} + aw_1 f_{n+1} + bw_0 f_{n+1} \\
 + \sum_{j=0}^n (aw_{n+2-j} + bw_{n+1-j} + cw_{n-j}) f_j \\
 = a f_{n+1}.
 \end{aligned}$$

□

By linearity, with the notation of Propositions 8.3 and 8.4 we conclude

8.5 Theorem. *The solutions of the linear second order recurrence*

$$a x_{n+2} + b x_{n+1} + c x_n = f_{n+1}$$

are given by the two-parameter family of sequences

$$x_n = c_1 u_n + c_2 v_n + \frac{1}{a} (w * f)_n, \quad c_1, c_2 \in \mathbb{R},$$

where $\{u_n\}$ and $\{v_n\}$ are defined in (8.5) and $\{w_n\}$ in (8.8).

d. \mathcal{Z} -transform and Laplace transform

Linear difference equations can be solved also using the method of generating functions or, better, a slight modification of it known as the \mathcal{Z} -transform, see 8.7 below.

Let $a = \{a_n\}$ be a sequence of complex numbers which grows at most exponentially, $|a_n| \leq CM^n$ for some $M > 0$. The \mathcal{Z} -transform of $\{a_n\}$ is the complex-valued function

$$\mathcal{Z}\{a\}(z) := \sum_{n=0}^{\infty} a_n \frac{1}{z^n}$$

that is defined at least in $\{z \mid |z| > M\}$. Of course

$$\mathcal{Z}\{a\}(z) = \mathcal{T}\{a\}\left(\frac{1}{z}\right),$$

$\mathcal{T}(a)$ being the generating function of $a = \{a_n\}$. Using properties of power series, we see that

- (i) $\mathcal{Z}\{a\}$ uniquely determines $\{a_n\}$,
- (ii) \mathcal{Z} is linear, i.e., if $\lambda, \mu \in \mathbb{C}$ and $a = \{a_n\}$, $b = \{b_n\}$ grow at most exponentially, then

$$\mathcal{Z}\{\lambda a + \mu b\}(z) = \lambda \mathcal{Z}\{a\}(z) + \mu \mathcal{Z}\{b\}(z), \quad |z| \text{ large.}$$

(iii) If $\mathbf{e}_k := \{\underbrace{(0, \dots, 0)}_k, 1, 0, 0, \dots\}$ is the *Kronecker sequence* then

$$\mathcal{Z}\{\mathbf{e}_k\}(z) := \frac{1}{z^k}.$$

(iv) If $a = \{a_n\}$, and

$$\tau_k\{a\} := \{\underbrace{(0, \dots, 0)}_k, a_0, a_1, a_2, \dots\}$$

is the *forward shift* by k places, then

$$\mathcal{Z}\{\tau_k\{a\}\}(z) = \sum_{n=k}^{\infty} a_n \frac{1}{z^{n+k}} = \frac{1}{z^k} \mathcal{Z}\{a\}(z).$$

(v) If $a = \{a_n\}$, and $\tau_{-k}\{a\} := \{a_{n+k}\}_n$ is the *backward shift* by k places, then

$$\mathcal{Z}\{\tau_{-k}\{a\}\}(z) = \sum_{n=0}^{\infty} a_{n+k} \frac{1}{z^n} = z^k \left(\mathcal{Z}\{a\}(z) - a_0 - \frac{a_1}{z} - \frac{a_2}{z^2} - \dots - \frac{a_{k-1}}{z^{k-1}} \right).$$

(vi) \mathcal{Z} transforms the convolution product of sequences into the product of the transformed functions, see Theorem 7.34,

$$\begin{aligned} \mathcal{Z}\{a * b\}(z) &= \sum_{n=0}^{\infty} (a * b)_n \frac{1}{z^n} = \left(\sum_{n=0}^{\infty} a_n \frac{1}{z^n} \right) \left(\sum_{n=0}^{\infty} b_n \frac{1}{z^n} \right) \\ &= \mathcal{Z}\{a\}(z) \mathcal{Z}\{b\}(z). \end{aligned}$$

The notion of \mathcal{Z} -transform (and of generating function) is very useful in several fields: in combinatorics, as we have seen, in probability, in data sampling, in the study of digital filters, just to mention a few. The \mathcal{Z} -transform is known, especially to engineers, as the *discrete* version of the *Laplace transform*, which is particularly useful when studying the Cauchy problem for linear ODE.

The *Laplace transform* of a continuous function that grows at infinity less than exponentially is defined by

$$\mathcal{L}\{f\}(z) := \int_0^{\infty} f(t) e^{-zt} dt, \quad \Re z > 0.$$

If f is the piecewise constant function defined by $f(t) = a_n$ if $n \leq t < n+1$, then

$$\mathcal{L}\{f\}(z) = \frac{e^{-z} - 1}{-z} \sum_{n=0}^{\infty} a_n e^{-nz} = \frac{1 - e^{-z}}{z} \mathcal{Z}\{a\}(e^z).$$

Also, if for all $h \in \mathbb{N}$ we set

$$f_h(t) := \begin{cases} a_n & \text{if } n < t < n + \frac{1}{h}, \\ 0 & \text{otherwise,} \end{cases}$$

then

$$\mathcal{L}\left\{\frac{1}{h}f_h\right\}(z) = \frac{1 - e^{-hz}}{hz} \mathcal{Z}\{a\}(e^z);$$

therefore

$$\lim_{h \rightarrow 0^+} \mathcal{L}\left\{\frac{1}{h}f_h\right\} = \mathcal{Z}\{a\}(e^z).$$

The functions $(1/h)f_h$ may be thought of as approximations of “impulses” concentrated at the integers.

e. Fibonacci's numbers

The simplest possible recurrence in which each number depends on the previous two is the one defining Fibonacci's numbers. They occur in a wide variety of situations. Here is how Leonardo Pisano (1170–1250), called Fibonacci, came to them.

Assume that every month every couple of rabbits gives birth to a couple of rabbits that can reproduce from their second month of life on. How many couples of rabbits are there after n months if we start with a newborn couple? If $\{f_n\}$ is such a number, of course, $f_1 = f_2 = 1$, moreover with a newborn living at the n -th month, f_n , are the ones of the previous month plus those generated by the rabbits who were alive two months before,

$$f_{n+2} = f_{n+1} + f_n.$$

Fibonacci's sequence $\{f_n\}$ is defined by

$$\begin{cases} f_0 = 0, & f_1 = 1, \\ f_{n+2} = f_{n+1} + f_n, & n \geq 0. \end{cases}$$

If we write the characteristic equation

$$z^2 - z - 1 = 0$$

of the recurrence to get its solutions, $\lambda = \frac{1+\sqrt{5}}{2}$ and $\mu = -\frac{\sqrt{5}-1}{2}$, we conclude on account of Section 8.1.1 that

$$f_n = c_1\lambda^n + c_2\mu^n \quad n \geq 0,$$

where c_1, c_2 solve

$$\begin{cases} f_0 = c_1 + c_2 = 0, \\ f_1 = c_1\lambda_1 + c_2\lambda_2 = 1. \end{cases}$$

Therefore

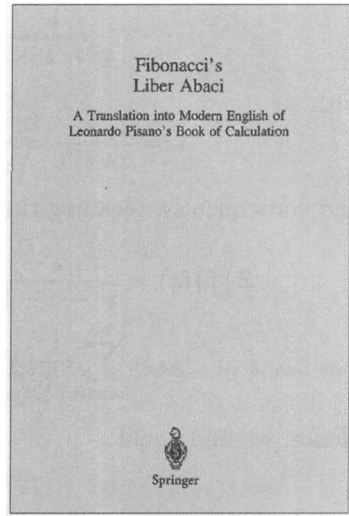
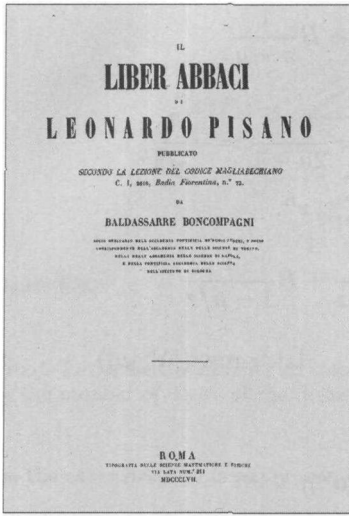


Figure 8.3. Frontispieces of the first printed edition and the first English translation of *Liber Abaci* of Leonardo Pisano (1170–1250), called Fibonacci.

8.6 Proposition (Binet's formula). We have

$$f_n = \frac{1}{\sqrt{5}} (\lambda^n - \mu^n) \quad \forall n \geq 0.$$

Since $|\mu|^n/\sqrt{5} \in]0, 1/2[$, and μ is negative if n is odd and positive if n is even, f_n is the integer part of $\lambda^n/\sqrt{5}$ if n is odd, and the integer part of $\lambda^n/\sqrt{5}$ plus 1 if n is even. In any case, f_n is the closest integer to $\lambda^n/\sqrt{5}$.

8.7 Fibonacci's numbers by \mathcal{Z} -transform. One can solve the Fibonacci recurrence also using the \mathcal{Z} -transform. In fact, multiplying the n -th recurrence relation by $\frac{1}{z^n}$ and summing, we get

$$z^2 \sum_{n=0}^{\infty} f_{n+2} \frac{1}{z^{n+2}} - z \sum_{n=0}^{\infty} f_{n+1} \frac{1}{z^{n+1}} - \sum_{n=0}^{\infty} f_n \frac{1}{z^n} = 0,$$

that is

$$z^2 \left(\mathcal{Z}\{f\}(z) - f_0 - \frac{f_1}{z} \right) - z \left(\mathcal{Z}\{f\}(z) - f_0 \right) - c \mathcal{Z}\{f\}(z) = 0,$$

i.e.,

$$\mathcal{Z}f(z) = \frac{z}{z^2 - z - 1}.$$

Notice that the denominator of $\mathcal{Z}\{f\}(z)$ is the characteristic equation of the Fibonacci recurrence. By Hermite's decomposition formula

$$\frac{1}{z^2 - z - 1} = A \frac{1}{z - \lambda} + B \frac{1}{z - \mu}$$

with

$$A := \frac{1}{2\lambda - 1} = \frac{1}{\sqrt{5}}, \quad B := \frac{1}{2\mu - 1} = -\frac{1}{\sqrt{5}}$$

and consequently, recalling that $\frac{1}{1-z} = \sum_{n=0}^{\infty} z^n$,

$$\begin{aligned} \mathcal{Z}\{f\}(z) &= \frac{z}{z^2 - z - 1} = A \frac{1}{1 - \lambda/z} + B \frac{1}{1 - \mu/z} \\ &= \sum_{n=0}^{\infty} \frac{1}{\sqrt{5}} (\lambda^n - \mu^n) \frac{1}{z^n}, \quad |z| > \max(|\lambda|, |\mu|). \end{aligned}$$

Hence, we find again

$$f_n = \frac{1}{\sqrt{5}} (\lambda^n - \mu^n).$$

In terms of Fibonacci numbers one can give a sharp estimate of the number of steps needed to end Euclid's algorithm.

8.8 Proposition. *Let $a, b \in \mathbb{N}$ with $0 < b \leq a$. If Euclid's algorithm on a and b ends in n steps, then $a \geq f_{n+2}$ and $b \geq f_{n+1}$.*

In other words, if $b < f_{n+1}$ or $a < f_{n+2}$, then Euclid's algorithm ends in at most $n - 1$ steps.

Proof. Write Euclid's algorithm as

$$\begin{cases} r_{-1} = a, & r_0 = b, \\ r_{j+1} = r_{j-1} - q_j r_j \end{cases}$$

until $r_{n+1} = 0$, so that Euclid's algorithm has n steps. Observe that

$$r_{n+1-j} \geq f_j \quad \forall j \in \{-1, 0, \dots, n+1\}.$$

Since we have $r_{n+1} = 0 = f_0$, $r_n = \text{g.c.d.}(a, b) \geq 1 = f_1$, and by induction

$$r_{n+1-j} = q_{n-j} r_{n-j} + r_{n-j-1} \geq f_{j-1} + f_{j-2} = f_j,$$

we infer

$$a = r_{-1} \geq f_{n+2}, \quad \text{and} \quad b = r_0 \geq f_{n+1}.$$

□

Notice that the estimate on the number of steps of Euclid's algorithm in Proposition 8.8 is sharp, since for $a = f_{n+2}$ and $b = f_{n+1}$, we have $r_1 = f_n$, $r_2 = f_{n-1}$, $\text{g.c.d.}(f_{n+2}, f_{n+1}) = r_n = f_1 = 1$, $r_{n+1} = f_0 = 0$. Thus Euclid's algorithm stops in n steps.

8.9 Corollary (Lamé). *The number of steps needed to end Euclid's algorithm does not exceed five times the number of the (decimal) digits of the divisor.*

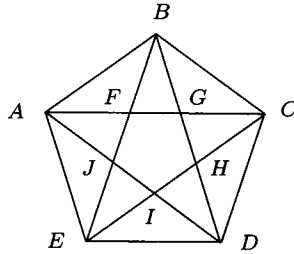


Figure 8.4.

Proof. Let n be the number of steps of Euclid's algorithm to divide a by b , and let k be the number of digits of the divisor. From Proposition 8.8,

$$10^k > b \geq f_{n+1}.$$

On the other hand, it is easily seen by induction that

$$f_{n+1} \geq \left(\frac{8}{5}\right)^{n-1}.$$

Since $(8/5)^5 > 10$ we have

$$10^{5k} \geq \left(\frac{8}{5}\right)^{5(n-1)} > 10^{n-1},$$

that is $n - 1 \leq 5k$. □

8.10 ¶. The number $\tau = \frac{1+\sqrt{5}}{2}$ is the *golden ratio*⁷ of Greek geometers. With reference to Figure 8.4, show that, if 1 is the side, then

- (i) the length of the diagonal is the golden ratio τ ,
- (ii) the side of the internal pentagon is τ^{-2} .

8.1.2 Some nonlinear examples

a. Simple examples

8.11 Example. Consider the recurrence

$$\begin{cases} x_0 = \alpha > 0, \\ x_{n+1} = \sqrt{x_n}, \quad n \geq 0. \end{cases} \quad (8.9)$$

If $\alpha = 1$, then $x_n = 1 \forall n$. If $\alpha > 1$, we see by induction that $x_n > 1 \forall n$, hence $x_{n+1} = \sqrt{x_n} \leq x_n, \forall n$, i.e., $\{x_n\}$ is decreasing, therefore $x_n \rightarrow L$ and $1 \leq L \leq \alpha$; actually, passing to the limit in (8.9), we see that $L = \sqrt{L}$, i.e., $L = 1$. Similarly, if $\alpha < 1$, then $x_n < 1 \forall n$, $\{x_n\}$ is increasing and $x_n \rightarrow 1$. We conclude that for any $\alpha > 0$, the sequence $\{x_n\}$ defined by (8.9) converges to 1.

Alternatively, it is easily guessed and proved that the sequence defined by (8.9) is $x_n = \alpha^{2^{-n}}, n \geq 0$, thus $x_n \rightarrow 1$.

⁷ The *golden ratio* is the inverse of the *golden mean*, which is the proportion of the division of a segment so that the smaller is to the larger as the larger is to the whole.

8.12 Example. Consider the recurrence

$$\begin{cases} x_0 = \alpha > 0, \\ x_{n+1} = \frac{1}{\sqrt{x_n}}, \end{cases} \quad n \geq 0. \quad (8.10)$$

If $\alpha = 1$, then $x_n = 1 \forall n$. Also, if $\alpha > 1$ we have $x_n > 1$ if n is even and $x_n < 1$ if n is odd. Moreover

$$x_{2n+2} = \frac{1}{\sqrt{x_{2n+1}}} = \sqrt[4]{x_{2n}}.$$

We deduce $x_{2n} = \alpha^{4^{-n}}$, $n \geq 0$, hence

$$x_{2n+1} = \left(\frac{1}{\sqrt{\alpha}} \right)^{4^{-n}},$$

concluding that $x_{2n}, x_{2n+1} \rightarrow 1$ and $x_n \rightarrow 1$ since even and odd integers exhaust all integers.

8.13 Example. A limit situation occurs for the recurrence

$$\begin{cases} x_0 = \alpha > 0, \\ x_{n+1} = \frac{1}{x_n} \end{cases} \quad n \geq 0.$$

Clearly $x_n = \alpha$ if n is even and $x_n = 1/\alpha$ if n is odd. We conclude that $\{x_n\}$ has limit if and only if $\alpha = 1$.

b. Evaluating algorithm performance

In evaluating the performance of algorithms one considers a characteristic time as a function $T(n)$ of a parameter n which describes the size of data on which the algorithm works. Often, due to the structure of the algorithm, one gets recursive estimates on $T(n)$ of the type⁸

$$T(2n) \leq 2T(n) + n, \quad \forall n.$$

One can prove that in fact this estimate is equivalent to the estimate $T(n) \leq Cn \log n$, i.e., using the *Landau notation*, to

$$T(n) = O(n \log n).$$

8.14 Proposition. Let $T : \mathbb{N} \rightarrow \mathbb{R}$ be a positive increasing function such that

$$T(\tau n) \leq \tau^\alpha T(n) + Bn^\beta, \quad \forall n \geq 0, \quad (8.11)$$

where $\tau \in \mathbb{N}$, $\tau \geq 2$, $B > 0$, $\alpha > 0$ and $\beta > 0$ are independent of n . Then

- (i) if $\alpha \neq \beta$, then $T(n) = O(n^{\max(\alpha, \beta)})$, i.e., there exists a constant $C = C(\alpha, \beta, \tau, T(1), B)$ such that $T(n) \leq C n^{\max(\alpha, \beta)} \forall n$,
- (ii) if $\alpha = \beta$, then $T(n) = O(n^\alpha \log n)$, i.e., $T(n) \leq C n^\alpha \log n \forall n \geq 2$ for a suitable constant C depending on $T(1), B, \alpha$ and τ .

⁸ See e.g., A. Aho, J. H. Hopcroft and J. D. Ullman, *Data Structures and Algorithms*, Addison-Wesley, 1983.

Proof. We deduce from (8.11)

$$\begin{aligned} T(\tau) &\leq \tau^\alpha T(1) + B, \\ T(\tau^2) &\leq \tau^\alpha T(\tau) + B\tau^\beta \leq \tau^{2\alpha} T(1) + B\tau^\alpha + B\tau^\beta, \quad \dots \end{aligned}$$

and inductively

$$T(\tau^{k+1}) \leq \tau^{(k+1)\alpha} T(1) + B\tau^{k\beta} \sum_{j=0}^k \tau^{j(\alpha-\beta)}, \quad \forall k. \quad (8.12)$$

(i) If $\alpha < \beta$, (8.12) yields

$$T(\tau^{k+1}) \leq \tau^{(k+1)\alpha} T(1) + B\tau^{k\beta} \frac{\tau^{(\alpha-\beta)(k+1)-1}}{\tau^{\alpha-\beta}-1} \leq C\tau^{k\beta}$$

where $C := \tau T(1) + B \frac{\tau}{\tau^{\alpha-\beta}-1}$.

Given $n \in \mathbb{N}$, we can choose k in such a way that $\tau^k \leq n < \tau^{k+1}$, and conclude

$$T(n) \leq T(\tau^{k+1}) \leq C\tau^{k\beta} \leq Cn^\beta,$$

since $T(n)$ is increasing. Similarly, if $\alpha > \beta$, (8.12) yields

$$T(\tau^{k+1}) \leq \tau^{(k+1)\alpha} T(1) + B\tau^{k\beta} \frac{\tau^{(\alpha-\beta)(k+1)-1}}{\tau^{\alpha-\beta}-1} \leq C\tau^{k\alpha}$$

where $C := \tau T(1) + B \frac{\tau}{\tau^{\alpha-\beta}-1}$. Thus (i) is proved.

(ii) If $\alpha = \beta$, (8.12) yields

$$T(\tau^{k+1}) \leq \tau^{(k+1)\alpha} T(1) + B\tau^{(k\alpha)}(n+1) \leq \tau^{k^\alpha} (\tau T(1) + B(n+1)),$$

hence, if k is chosen in such a way that $\tau^k \leq n < \tau^{k+1}$, i.e., $k \leq \log_\tau n \leq k+1$, we find

$$T(n) \leq T(\tau^{k+1}) \leq \tau^{k^\alpha} (\tau^\alpha T(1) + B(\log_\tau n + 1)) \leq n^\alpha (\tau^\alpha + B + B \log_\tau n),$$

$T(n)$ being increasing, hence,

$$T(n) \leq C n^\alpha \log n \quad \forall n \geq 2,$$

for a suitable constant C . □

8.15 QuickSort. The average number of comparison steps C_n made by the QuickSort algorithm for sorting n random elements is given by

$$C_0 = 0, \quad C_n = n + 1 + \frac{2}{n} \sum_{j=0}^{n-1} C_j, \quad n \geq 1.$$

Multiplying by n we get for $n \geq 1$,

$$\begin{cases} nC_n = n^2 + n + 2 \sum_{j=0}^{n-1} C_j, \\ (n+1)C_{n+1} = (n+1)^2 + (n+1) + 2 \sum_{j=0}^n C_j, \end{cases}$$

and subtracting the first term from the second

$$C_{n+1} = \frac{n+2}{n+1} C_n + 2.$$

This is written, for $s_n := C_n/(n+1)$, as

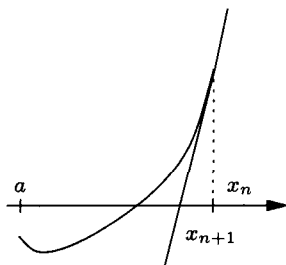


Figure 8.5.

$$\begin{cases} s_0 = 0, \\ s_{n+1} = s_n + 2/(n+2), \end{cases}$$

or $s_n := \sum_{j=0}^n \frac{2}{j+1} - 2 = \sum_{j=1}^n 2/(j+1)$. According to the above

$$C_n := (n+1)s_n = 2(n+1) \sum_{j=1}^n \frac{1}{j+1} = 2(n+1)H_{n+1} - 1$$

where $H_n := \sum_{j=1}^n 1/j$ is the n -th partial harmonic sum. Consequently (6.21) yields

$$C_n \sim 2(n+1) \log(n+1) \quad \forall n \quad \text{or} \quad C_n = O(n \log n).$$

c. Rate of convergence

8.16 Newton's approximation method. Let $f : [a, b] \rightarrow \mathbb{R}$ be a continuous function with $f(a) < 0$ and $f(b) > 0$. Theorem 2.51 states that f has at least one zero, and the proof provides an algorithm to approximate that zero. In the case in which f is also convex or concave, *Newton's method* turns out to be more efficient.

Assume f convex, continuous in $[a, b]$, $f(a) < 0$, $f(b) > 0$ so that f has a unique zero $c \in [a, b]$, and (see, e.g., Chapter 4 of [GM1])

$$f(x) > 0, \quad f'(x) > f'(c) > 0 \quad \text{for } x \in [c, b]. \quad (8.13)$$

Let $\{x_n\}$ be the sequence defined by the recurrence

$$\begin{cases} x_0 = b, \\ x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \end{cases}$$

Clearly x_{n+1} is the point where the tangent to the graph of f at the point $(x_n, f(x_n))$ intersects the x -axis. Since f is convex, $c \leq x_n \forall n$ and by (8.13), $x_{n+1} \leq x_n \forall n$. Thus the sequence $\{x_n\}$ converges to a point $L \in [c, b]$ that, as we see passing to the limit in the recurrence, is given by

$$L = L - \frac{f(L)}{f'(L)}, \quad \text{i.e.,} \quad f(L) = 0, \quad \text{i.e.,} \quad L = c.$$

$\{x_n\}$ is therefore a sequence of approximants of c from above. A sequence of approximations $\{y_n\}$ from below can be obtained by defining $y_0 = a$ and y_{n+1} as the point at which the secant through $(y_n, f(y_n))$ and $(x_n, f(x_n))$ intersects the x -axis.

Notice that Heron's algorithm in Exercise 2.99 is Newton's method applied to $f(x) = x^2 - \alpha$.

Let $g : [a, b] \rightarrow [a, b]$ be a function of class $C^2([a, b])$ and let $\{x_n\}$ be defined by the recurrence

$$\begin{cases} x_0 := \alpha \in [a, b], \\ x_{n+1} = g(x_n). \end{cases} \quad (8.14)$$

If $\{x_n\}$ converges, then $x_n \rightarrow L \in [a, b]$ with $L = g(L)$, that is L is a *fixed point* of g . From Taylor's formula with Lagrange remainder (see, e.g., [GM1]),

$$x_{n+1} = g(x_n) = g(L) + g'(\xi_n)(x_n - L), \quad \xi = \xi(y) \in [y, L],$$

we infer for the error $\delta_n := |x_n - L|$,

$$\delta_{n+1} = |x_{n+1} - L| = |g(x_n) - L| \leq M|x_n - L| = M\delta_n$$

where $M := \sup_{x \in [a, b]} |g'(x)|$. Since we assumed $\delta_n \rightarrow 0$, we have $\delta_n < (1/2)^n$ and $\{\delta_n\}$ decays exponentially to 0.

If moreover $g'(L) = 0$, Taylor's formula with Lagrange remainder yields also

$$g(x_n) = g(L) + \frac{g''(\eta_n)}{2}(x_n - L)^2,$$

hence

$$\delta_{n+1} = |x_{n+1} - L| = |g(x_n) - g(L)| \leq N|x_n - L|^2 = N\delta_n^2$$

where $N := \sup_{x \in [a, b]} |g''(x)|$. Therefore, we find for $p \in \mathbb{N}$ and $n \geq 1$,

$$\delta_{n+p} \leq \frac{1}{M} (M\delta_p)^{2^n}. \quad (8.15)$$

We say that a sequence $\{x_n\}$ *converges rapidly* to $L \in \mathbb{R}$ if $|x_n - L| < a^{2^n}$ definitively, with $a < 1$. From the previous argument we then conclude the following.

8.17 Proposition. *Let $g \in C^2([a, b])$, let $L \in [a, b]$ be a fixed point of g , $g(L) = L$, and let $\{x_n\}$ be the sequence defined by (8.14). If $\liminf_{n \rightarrow \infty} |x_n - L| = 0$, (in particular if $\{x_n\}$ converges to L), then $\{x_n\}$ converges rapidly to L .*

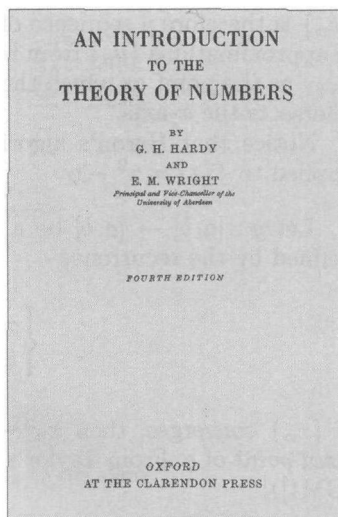


Figure 8.6. A classical introduction to number theory.

In the case of Newton's approximating sequence

$$\begin{cases} x_0 = \alpha \in [a, b], \\ x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \end{cases}$$

if $\{x_n\}$ converges, then the limit L is a fixed point of the function

$$g(y) := y - \frac{f(y)}{f'(y)}, \quad y \in [a, b].$$

Assuming moreover that $f \in C^3([a, b])$ and $f'(L) \neq 0$. Then $g \in C^2([a, b])$ and $g'(L) = 0$. Therefore, if $\{x_n\}$ converges to L , then $\{x_n\}$ converges rapidly to L .

8.18 ¶. Let $\{x_n\}$ be a sequence of positive real numbers such that

$$x_{n+1} \leq C B^n x_n^{1+\epsilon}$$

with $C > 0$, $B > 1$ and $\epsilon > 0$. If $x_0 \leq C^{-1/\epsilon} B^{-1/\epsilon^2}$, then $x_n \leq B^{-n/\epsilon} x_0$, hence $x_n \rightarrow 0$.

8.1.3 Continued fractions

a. Definitions and elementary properties

The *finite continued fraction operation* consists in computing, starting from $n+1$ nonnegative numbers $\{a_0, \dots, a_n\}$, which are all positive except for a_0 , the number

$$a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \dots + \frac{1}{a_n}}}}$$

One refers to the result as to the (finite) *continued fraction* of a_0, \dots, a_n . Since the previous notation is heavy, one prefers lighter notation, as

$$a_0 + \frac{1}{a_1 +} \frac{1}{a_2 +} \dots \frac{1}{a_{n-1} +} \frac{1}{a_n},$$

or, as we shall do,

$$[a_0, \dots, a_n].$$

The numbers a_0, \dots, a_n are called the *quotients* of the continued fraction $[a_0, \dots, a_n]$, while for $0 \leq k \leq n$, the continued fraction $[a_0, \dots, a_k]$ is called the *k-th convergent* of $[a_0, \dots, a_n]$.

Observe that $[a_0] = a_0$,⁹ $[a_0, a_1] = a_0 + \frac{1}{a_1}$, and more generally

$$\begin{aligned} [a_0, \dots, a_n] &= [a_0, \dots, a_{n-2}, a_{n-1} + \frac{1}{a_n}] = [a_0, \dots, a_{n-2}, [a_{n-1}, a_n]], \\ [a_0, \dots, a_n] &= a_0 + \frac{1}{[a_1, \dots, a_n]} = [a_0, [a_1, \dots, a_n]], \\ [a_0, \dots, a_n] &= [a_0, \dots, a_k, [a_{k+1}, \dots, a_n]] \quad \forall k, 0 \leq k \leq n. \end{aligned} \tag{8.16}$$

Finally, observe that the map

$$(a_0, \dots, a_n) \mapsto [a_0, \dots, a_n]$$

is strictly increasing in each of the variables with an even index and strictly decreasing in each of the variables with an odd index.

Computing $[a_0, \dots, a_n]$ by its definition consists in the following: start from the last a_n , take the inverse $1/a_n$, add a_{n-1} , compute the inverse of the result, add a_{n-2} and so on by downward induction until one adds a_0 . The following iterative scheme,

$$[a_0] = \frac{a_0}{1}, \quad [a_0, a_1] = \frac{a_0 a_1 + 1}{a_1}, \quad [a_0, a_1, a_2] = \frac{a_0 a_1 a_2 + a_2 + a_0}{a_2 a_1 + 1},$$

computes $[a_0, \dots, a_n]$ by upward induction, and reduces the computation of $[a_0, \dots, a_n]$ to a sum.

Let $[a_0, \dots, a_n]$ be a continued fraction. Define p_0, \dots, p_n , and q_0, \dots, q_n by

$$\begin{cases} p_0 = a_0, \\ q_0 = 1, \end{cases} \quad \begin{cases} p_1 = a_0 a_1 + 1, \\ q_1 = a_1, \end{cases} \tag{8.17}$$

⁹ In this context $[a_0] = a_0$ and not, as usual, the integral part of a_0 . We denote instead the floor of x by $x//1$, $//$ being the integral division.

and, for $k = 2, \dots, n$,

$$\begin{cases} p_k = a_k p_{k-1} + p_{k-2}, \\ q_k = a_k q_{k-1} + q_{k-2}. \end{cases} \quad (8.18)$$

Then $q_k \geq 0 \forall k$ and we have the following.

8.19 Proposition. *We have*

$$\frac{p_k}{q_k} = [a_0, \dots, a_k], \quad \forall k = 0, 1, \dots, n. \quad (8.19)$$

Moreover

$$p_k q_{k-1} - p_{k-1} q_k = (-1)^{k-1}, \quad k = 1, \dots, n \quad (8.20)$$

$$p_k q_{k-2} - p_{k-2} q_k = (-1)^k a_k, \quad k = 2, \dots, n. \quad (8.21)$$

In particular

$$[a_0, \dots, a_k] = \frac{p_0}{q_0} + \sum_{j=0}^{k-1} \frac{(-1)^j}{q_j q_{j+1}}, \quad k = 1, \dots, n. \quad (8.22)$$

Proof. We first prove (8.19) by induction on k . Of course, (8.19) holds true for $k = 0, 1$. By induction assume the claim true for k , then

$$\begin{aligned} [a_0, a_1, \dots, a_{k+1}] &= \left[a_0, a_1, \dots, a_{k-1}, a_k + \frac{1}{a_{k+1}} \right] = \frac{(a_k + \frac{1}{a_{k+1}})p_{k-1} + p_{k-2}}{(a_k + \frac{1}{a_{k+1}})q_{k-1} + q_{k-2}} \\ &= \frac{a_{k+1}(a_k p_{k-1} + p_{k-2}) + p_{k-1}}{a_{k+1}(a_k q_{k-1} + q_{k-2}) + q_{k-1}} = \frac{a_{k+1}p_k + p_{k-1}}{a_{k+1}q_k + q_{k-1}} = \frac{p_{k+1}}{q_{k+1}}. \end{aligned}$$

Then (8.19) follows. As

$$\begin{aligned} p_k q_{k-1} - p_{k-1} q_k &= (a_k p_{k-1} + p_{k-2})q_{k-1} - p_{k-1}(a_k q_{k-1} + q_{k-2}) \\ &= -(p_{k-1} q_{k-2} - p_{k-2} q_{k-1}), \end{aligned}$$

by repeating the argument, we get

$$p_k q_{k-1} - p_{k-1} q_k = (-1)^{k+1} (p_1 q_0 - p_0 q_1) = (-1)^{k-1},$$

i.e., (8.20). Also

$$\begin{aligned} p_k q_{k-2} - p_{k-2} q_k &= (a_k p_{k-1} + p_{k-2})q_{k-2} - p_{k-2}(a_k q_{k-1} + q_{k-2}) \\ &= a_k (p_{k-1} q_{k-2} - p_{k-2} q_{k-1}) = (-1)^k a_k. \end{aligned}$$

i.e., (8.21).

Finally, (8.20) is written as

n	p_n/q_n	p_n/q_n
0	0.0000000000000000	0/1
1	1.0000000000000000	1/1
2	0.6666666666666667	2/3
3	0.7500000000000000	3/4
4	0.748691099476440	143/191
5	0.748704663212435	289/386
6	0.748701973001038	721/963
7	0.748702422145329	1731/2312
1731/2312 = 0.748702422145329 = [0, 1, 2, 1, 47, 2, 2, 2]		

Figure 8.7. The continued fraction expansion of 1731/2312.

$$\frac{p_k}{q_k} - \frac{p_{k-1}}{q_{k-1}} = \frac{(-1)^{k-1}}{q_k q_{k-1}}$$

for $k = 1, \dots, n$, hence

$$\frac{p_k}{q_k} = \frac{p_0}{q_0} + \sum_{j=1}^k \frac{(-1)^{j-1}}{q_j q_{j-1}},$$

hence (8.22), by taking into account (8.19). \square

In the rest of this section we are interested in *simple continued fractions*, that is, in continued fractions in which all the quotients are integers. Clearly in this case the continued fraction is a rational number.

8.20 Definition. A continued fraction $[a_0, \dots, a_n]$ is simple if $a_i \in \mathbb{N} \forall i$ and $a_i \geq 1 \forall i \geq 1$.

For simple continued fractions, Proposition 8.19 is particularly useful. We have the following.

8.21 Proposition. Let $[a_0, \dots, a_n]$ be a simple continued fraction, and let $p_k/q_k := [a_0, \dots, a_k]$ be irreducible. Then

- (i) $\{p_k\}$ and $\{q_k\}$ are the numbers defined in (8.17), (8.18),
- (ii) $q_1 \geq q_0$ and $q_k > q_{k-1} \forall k \geq 2$,
- (iii) $q_k \geq k \forall k$, and $q_k > k \forall k \geq 3$.

Proof. The numbers p_0, \dots, p_k and q_0, \dots, q_k defined by (8.17), (8.18) are integers by definition, moreover they are coprime by (8.20) and $p_k/q_k = [a_0, \dots, a_k]$ by (8.19). Thus (i) follows. Finally we have $q_n = a_n q_{n-1} + q_{n-2} \geq q_{n-1} + 1$ if $n \geq 2$, hence (ii), while (iii) follows at once since $q_n \geq q_{n-1} + q_{n-2} \geq q_{n-1} + 1 \geq n$ if $n \geq 3$. \square

A continued fraction does not fix its quotients as, for instance,

$$\begin{aligned} [a_0, \dots, a_n] &= [a_0, \dots, a_{n-1}, a_n - 1, 1] & \text{if } a_n > 1, \\ [a_0, \dots, a_n] &= [a_0, \dots, a_{n-1} + 1] & \text{if } a_n = 1. \end{aligned}$$

However a *simple* continued fraction fixes its quotients apart from the previous ambiguity. More precisely, we have the following.

8.22 Proposition. *Let $[h_0, \dots, h_n]$ and $[a_0, \dots, a_m]$ be two simple continued fractions. Suppose that $[h_0, \dots, h_n] = [a_0, \dots, a_m]$ and, for convenience, $m \geq n$. Then*

- either $m = n$ and $h_i = a_i \forall i = 0, \dots, n$,
- or $m = n + 1$, $h_i = a_i \forall i = 1, \dots, n - 1$, $h_n = a_n + 1$ and $a_m = 1$.

Proof. We proceed by induction on n . Let $n = 0$. Either $m = 0$, hence $h_0 = [h_0] = [a_0] = a_0$, or $m > 0$. In this case we have

$$h_0 = [h_0] = [a_0, \dots, a_m] = a_0 + \frac{1}{[a_1, \dots, a_m]}$$

from which we infer $[a_1, \dots, a_m] = 1$, which in turn implies $m = 1$, $a_1 = [a_1, \dots, a_m] = 1$ and $h_0 = a_0 + 1$.

Assuming now the claim true for simple continued fractions with n quotients, let us prove it for a simple continued fraction with $n + 1$ quotients. Assuming $n \geq 1$, by (8.16), we have

$$[h_0, [h_1, \dots, h_n]] = [a_0, [a_1, \dots, a_m]]$$

hence by the inductive assumption $h_0 = a_0$ and $[h_1, \dots, h_n] = [a_1, \dots, a_m]$. Using again the inductive assumption, we reach the conclusion. \square

8.23 Corollary. *Let $[a_0, \dots, a_n]$ and $[b_0, \dots, b_m]$ be two simple continued fractions. Suppose that $a_n, b_m \geq 2$, and that $[a_0, \dots, a_n] = [b_0, \dots, b_m]$. Then $n = m$ and $a_i = b_i, \forall i = 0, \dots, n$.*

8.24 Definition. Let $\{a_n\}$, $n = 0, 1, \dots$ be a list of real numbers such that $a_0 \geq 0$ and $a_i > 0$. We refer to the sequence

$$\left\{ [a_0, \dots, a_n] \mid n = 0, 1, \dots \right\}$$

as to an infinite continued fraction, and we write it as $[a_0, \dots, a_n, \dots]$. For any integer n , the (finite) continued fraction $[a_0, \dots, a_n]$ is called the n -th convergent of $[a_0, \dots, a_n, \dots]$. If $[a_0, \dots, a_n] \rightarrow x \in \mathbb{R}$ as $n \rightarrow \infty$, we also write $x = [a_0, \dots, a_n, \dots]$.

b. Developments as continuous fractions

The following algorithm leads us to simple continued fractions. Let x be a real number. Let a_0 be its *integral part*, $a_0 := x//1$, and let $\alpha_1 := x - a_0$ be its *fractional part*, so that

$$x = a_0 + \alpha_1, \quad a_0 \in \mathbb{N}, \quad a_0 \geq 0, \quad \alpha_1 \in [0, 1[.$$

If $\alpha_1 \neq 0$, we can write

$$x = a_0 + \frac{1}{\frac{1}{\alpha_1}},$$

and, since $1/\alpha_1 > 1$, we can reiterate the procedure,

$$\begin{aligned} a_1 &:= 1/\alpha_1 // 1, & \alpha_2 &:= 1/\alpha_1 - a_1, \\ \frac{1}{\alpha_1} &= a_1 + \alpha_2 = a_1 + \frac{1}{\frac{1}{\alpha_2}}, & &= [a_0, a_1 + \alpha_2] \end{aligned}$$

to write

$$\begin{aligned} x = a_0 + \alpha_1 &= a_0 + \frac{1}{a_1 + \alpha_2} = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + \alpha_3}} = \dots \\ &= [a_0, a_1, \dots, a_k + \alpha_{k+1}] \end{aligned} \tag{8.23}$$

for $k = 1, 2, \dots$ as long as $\alpha_k > 0$. Thus the algorithm either ends at the first $k = n$ at which $\alpha_{k+1} = 0$, and in this case $x = [a_0, \dots, a_n]$, or eventually continues indefinitely.

All the a_i 's we find in this way are nonnegative integers and $a_i \geq 1 \forall i \geq 1$, thus the resulting continued fractions are simple. We refer to that algorithm as to the *continued fraction expansion algorithm* and to the resulting list of continued fractions when applying the algorithm to x as to the *continued fraction expansion of x* .

Finally, we recall that, since $(a_0, \dots, a_n) \rightarrow [a_0, \dots, a_n]$ is strictly increasing in the variables with even index and decreasing in the variables with odd index,

$$\begin{aligned} [a_0, \dots, a_{2k}] &\leq x = [a_0, \dots, a_{2k-1}, a_{2k} + \frac{1}{\alpha_{2k+1}}] \\ &= [a_0, \dots, a_{2k}, a_{2k+1} + \frac{1}{\alpha_{2k+2}}] \\ &\leq [a_0, \dots, a_{2k+1}] \end{aligned} \tag{8.24}$$

as far as the continued fraction expansion algorithm continues.

8.25 Euclid's algorithm. Let $x > 0$. Euclid's algorithm is a means to find iteratively, if it exists, a common submultiple of $x > 0$ and 1, see Section 1.1 and Figure 1.5, by

Start with $x \in \mathbb{R}$, then compute $\{a_k\}$ and α_k with

$$\alpha_0 := \frac{1}{x}, \quad \begin{cases} \beta_k := 1/\alpha_k, \\ a_k := \beta_k // 1, \\ \alpha_{k+1} := \beta_k - a_k \end{cases}$$

for $k = 0, 1, \dots$, as far as $\alpha_{k+1} \neq 0$.

By construction

$$x = a_0 + \alpha_0 = [a_0 + \alpha_0] = [a_0, a_1 + \alpha_1] = \dots = [a_0, a_1, \dots, a_k + \alpha_{k+1}] = \dots$$

for $k = 0, 1, \dots$, as far as the algorithm continues. Eventually the algorithm stops at the first $k =: n$ for which $\alpha_{k+1} = 0$. In this case we also have

$$x = [a_0, a_1, \dots, a_n]. \quad (8.25)$$

Figure 8.8. The continued fraction expansion algorithm.

$$\begin{cases} r_0 = x, \quad r_1 = 1, \\ q_j := (r_{j-1}/r_j) // 1, \\ r_{j-1} = q_j r_j + r_{j+1}, \end{cases} \quad (8.26)$$

for $j = 1, \dots$, as long as $r_j > 0$. We refer to it as to *Euclid's algorithm starting from $(x, 1)$* . The algorithm eventually stops at the first $k =: n$ for which $r_{n+1} = 0$. In this case $x = s r_n$, $1 = t r_n$, $s, t \in \mathbb{N}$, and $x = s/t$ is rational. Moreover, the last quotient q_n is larger than or equal to 2. If conversely x is rational, then Euclid's algorithm surely stops after a finite number of steps. In fact, if $x = p/q$, p, q coprime, the remainders are $1/q$ times the corresponding remainders of Euclid's algorithm starting with (p, q) , which form a list of strictly decreasing integers, see Section 3.1.1. Thus *Euclid's algorithm, starting with $(x, 1)$, i.e., (8.26), stops after a finite number of steps if and only if x is rational.*

The development of x as a continued fraction is a rewriting of Euclid's algorithm (8.26). We in fact have the following.

8.26 Theorem. *Let $\{a_j\}$, and $\{\alpha_j\}$ be the numbers produced by the continued fraction expansion algorithm of x , and let $\{q_j\}$, $\{r_j\}$ be respectively the quotients and the remainders of Euclid's algorithm (8.26) starting with $(x, 1)$. Then*

$$q_j = a_j, \quad r_{j+1} = \alpha_1 \alpha_2 \dots \alpha_j \alpha_{j+1}.$$

Thus the continued fraction algorithm starting with $x > 0$ produces

- *either the quotients of a finite simple continued fraction $[a_0, \dots, a_n]$ such that $x = [a_0, \dots, a_n]$, if x is rational; in this case, if $x = p/q$, p, q coprime, we have $x = [a_0, \dots, a_n]$ where n is the number of steps in Euclid's algorithm to compute g.c.d. (p, q) and $a_n \geq 2$;*
- *or an infinite continued fraction $[a_0, \dots, a_n, \dots]$ if x is irrational.*

n	p_n/q_n	p_n/q_n	$1/(q_n q_{n+1})$
1	3.000000000000000	3/1	1e + 00
2	2.666666666666667	8/3	3e - 01
3	2.750000000000000	11/4	8e - 02
5	2.718750000000000	87/32	4e - 03
7	2.718309859154930	193/71	4e - 04
9	2.718283582089552	1457/536	4e - 06
11	2.718281835205993	23225/8544	1e - 07
13	2.718281828735696	49171/18089	6e - 09
$e = 2.718281828459045 = [2, 1, 2, 1, 1, 4, 1, 1, 6, 1, 1, 8, 1, 1, 10, \dots]$			

Figure 8.9. The continued fraction expansion of Euler number e .

If x is rational, then the continued fraction expansion $[a_0, \dots, a_n]$ of x is the only simple continued fraction with $a_n \geq 2$ which equals x , see Corollary 8.23.

8.27 ¶. Write a detailed proof of Theorem 8.26.

c. Infinite continued fractions

From Propositions 8.19 and 8.21 we easily get the following.

8.28 Theorem. Let $[a_0, \dots, a_n, \dots]$ be a simple infinite continued fraction, and let p_n/q_n be the irreducible representation of the n -th convergent $[a_0, \dots, a_n]$. Then $\{p_n/q_n\}$ converges to $x \in \mathbb{R}$,

$$x = [a_0, \dots, a_n, \dots],$$

where

$$x := a_0 + \sum_{j=0}^{\infty} \frac{(-1)^j}{q_j q_{j+1}}.$$

Moreover,

(i) $x - p_n/q_n$ is positive if n is even and negative if n is odd, so that

$$\frac{p_{2n}}{q_{2n}} \leq x \leq \frac{p_{2n+1}}{q_{2n+1}}, \quad (8.27)$$

(ii) we have

$$\frac{1}{q_n(q_n + q_{n+1})} < \left| x - \frac{p_n}{q_n} \right| < \frac{1}{q_n q_{n+1}},$$

(iii) we have

$$q_n x - p_n = \frac{(-1)^n \delta_n}{q_{n+1}}$$

where $0 < \delta_n < 1$,

(iv) the differences

$$|q_n x - p_n|, \quad \left| x - \frac{p_n}{q_n} \right|$$

are strictly decreasing as n increases.

Proof. From Proposition 8.21 the sequence $\{q_j q_{j+1}\}$ diverges and is strictly monotone $\forall j$. Moreover (8.20) yields

$$[a_0, \dots, a_n] = \frac{p_n}{q_n} = a_0 + \sum_{j=0}^{n-1} \frac{(-1)^j}{q_j q_{j+1}}.$$

Since the series

$$\sum_{j=0}^{\infty} \frac{(-1)^j}{q_j q_{j+1}}$$

is alternating and the absolute values of the terms tend strictly to zero, the Leibniz test applies, hence $p_n/q_n \rightarrow x$, where

$$x = a_0 + \sum_{j=0}^{\infty} \frac{(-1)^j}{q_j q_{j+1}}, \quad (8.28)$$

and (i) and the estimate from above of $|x - p_n/q_n|$ in (ii) hold. The estimate from below in (ii) follows also from the Leibniz test since

$$\begin{aligned} \left| x - \frac{p_n}{q_n} \right| &\geq \frac{1}{q_n q_{n+1}} - \frac{1}{q_{n+1} q_{n+2}} = \frac{a_{n+1}}{q_n (a_{n+1} q_{n+1} + q_n)} \\ &\geq \frac{1}{q_n (q_{n+1} + q_n)}. \end{aligned}$$

(iii) follows from (ii).

(iv) From (ii) we infer

$$\begin{aligned} |q_{n+1} x - p_{n+1}| &< \frac{1}{q_{n+2}} = \frac{q_n}{q_n (a_{n+2} q_n + q_{n+1})} \\ &\leq q_n \frac{1}{q_n (q_n + q_{n+1})} < |q_n x - p_n| \end{aligned}$$

thus $\{|q_n x - p_n|\}$ is strictly decreasing. As a consequence $\{|x - p_n/q_n|\}$ is strictly decreasing, too. \square

The next proposition shows that a simple infinite continued fraction is completely identified by its limit.

n	p_n/q_n	p_n/q_n	$1/(q_n q_{n+1})$
1	2.000000000000000	2/1	1e + 00
2	1.500000000000000	3/2	5e - 01
3	1.666666666666667	5/3	2e - 01
5	1.625000000000000	13/8	3e - 02
7	1.619047619047619	34/21	4e - 03
9	1.618181818181818	89/55	5e - 04
11	1.618055555555556	233/144	8e - 05
13	1.618037135278515	610/377	1e - 05
$\frac{1+\sqrt{5}}{2} = 1.618033988749895 = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, \dots]$			

Figure 8.10. The continued fraction expansion of the golden ratio.

8.29 Proposition. Let $[a_0, \dots, a_n, \dots]$ and $[b_0, \dots, b_n, \dots]$ be two infinite simple continued fractions that have the same limit. Then $a_n = b_n \forall n$.

Proof. For $i = 0, 1, 2, \dots$, denote by x_i the real numbers

$$x_i := [a_i, \dots, a_n, \dots] := \lim_{n \rightarrow \infty} [a_i, \dots, a_n]$$

which exist by Theorem 8.28. We first observe that

$$1 \leq a_i < x_i < a_i + 1 \quad \forall i, \quad (8.29)$$

in fact, since an infinite continued fraction never ends, $x_i = a_i + \frac{1}{x_{i+1}}$ is strictly larger than $a_i \forall i$. Consequently $x_i = a_i + \frac{1}{x_{i+1}} < a_i + \frac{1}{a_{i+1}} < a_i + 1$. Then (8.29) yields

$$0 < \frac{1}{a_i + 1} < \frac{1}{x_i} < \frac{1}{a_i} \leq 1 \quad \forall i \geq 1.$$

In particular, $1/x_i < 1$. If now $[a_0, \dots, a_n, \dots] = [b_0, \dots, b_n, \dots]$, then

$$a_0 + \frac{1}{x_1} = b_0 + \frac{1}{y_1}.$$

Since $a_0, b_0 \in \mathbb{N}$ and $1/x_1, 1/y_1 \in]0, 1[$, we then infer $a_0 = b_0$ and

$$[a_1, \dots, a_n, \dots] = [b_1, \dots, b_n, \dots].$$

We then iterate the previous argument to get $a_1 = b_1, a_2 = b_2, \dots$ □

We therefore conclude the following.

8.30 Theorem. Let x be irrational. Then there is a unique simple continued fraction that converges to x : the continued fraction expansion of x .

Proof. Let $[a_0, \dots, a_n, \dots]$ be the continued fraction expansion of x . By Theorem 8.28, $[a_0, \dots, a_n, \dots]$ converges to $y \in \mathbb{R}$ and

$$[a_0, \dots, a_{2n}] \leq y \leq [a_0, \dots, a_{2n+1}].$$

Since also

$$[a_0, \dots, a_{2n}] \leq x \leq [a_0, \dots, a_{2n+1}]$$

by construction, we infer $y = x$ letting $n \rightarrow \infty$. The uniqueness is stated as Proposition 8.29 \square

d. Irrationals and approximations by rationals

Actually the n -th convergent $p_n/q_n = [a_0, \dots, a_n]$ of the continued fraction expansion of x is the best approximation of x among all fractions whose denominator does not exceed q_n .

8.31 Theorem (Best rational approximations). *Let x be irrational and let p_n/q_n be the n -convergent of the continued fraction development of x . Then $\forall n \geq 2$, $\forall p, q \in \mathbb{N}$ coprime with $q \leq q_n$ and $p/q \neq p_n/q_n$, we have*

$$|q_n x - p_n| < |q x - p|, \quad \left| x - \frac{p_n}{q_n} \right| < \left| x - \frac{p}{q} \right|.$$

Proof. We have already proved that

$$|q_n x - p_n| < |q_{n-1} x - p_{n-1}|, \forall n \geq 0,$$

hence the claim follows by downward induction if we prove it for p, q such that $q_{n-1} < q \leq q_n$ and $p/q \neq p_n/q_n$.

If $q = q_n$, we have $|p - p_n| \geq 1$ and $q_{n+1} \geq 2$, hence

$$|q_n x - p_n| \leq \frac{1}{q_{n+1}} \leq \frac{1}{2} \leq \frac{1}{2} |p_n - p| \leq \frac{1}{2} |q_n x - p_n| + \frac{1}{2} |q x - p|.$$

If $q_{n-1} < q < q_n$, and $p/q \neq p_n/q_n$ we also have $p/q \neq p_{n-1}/q_{n-1}$. Write

$$\begin{cases} p = \alpha p_{n-1} + \beta p_n, \\ q = \alpha q_{n-1} + \beta q_n, \end{cases}$$

that is

$$\begin{cases} \alpha(p_n q_{n-1} - q_n p_{n-1}) = p q_n - q p_n, \\ \beta(p_n q_{n-1} - q_n p_{n-1}) = p q_{n-1} - q p_{n-1}, \end{cases}$$

i.e.,

$$\alpha = (-1)^{n-1} (p q_n - q p_n), \quad \beta = (-1)^{n-1} (p q_{n-1} - q p_{n-1}),$$

in particular α and β are nonzero integers. Since $q_{n-1} < q = \alpha q_{n-1} + \beta q_n \leq q_n$, α and β must have opposite signs, while $p_n - q_n x$ and $p_{n-1} - q_{n-1} x$ do have opposite signs. Thus

$$\alpha(p_{n-1} - q_{n-1}x) \quad \text{and} \quad \beta(p_n - q_nx)$$

have the same sign; hence

$$|p - qx| = |\alpha(p_{n-1} - q_{n-1}x)| + |\beta(p_n - q_nx)| \geq |p_{n-1} - q_{n-1}x| > |p_n - q_nx|.$$

□

8.32 Example. According to the definition of the continued fraction expansion algorithm, one easily sees that $[1, 1, \dots, 1, \dots]$ is the continued fraction expansion of the golden ratio, $\frac{1+\sqrt{5}}{2}$. Moreover $p_n = f_{n-1}$, $q_n = f_n$, $\{f_n\}$ being the sequence of Fibonacci numbers. Since in this case

$$\begin{cases} q_0 = 1, & q_1 = 1, \\ q_{n+2} = q_{n+1} + q_n, \end{cases} \quad \begin{cases} p_0 = 1, & p_1 = 2, \\ p_{n+2} = p_{n+1} + p_n, \end{cases}$$

we deduce

$$\frac{f_{n-1}}{f_n} \rightarrow \frac{1+\sqrt{5}}{2}, \quad \frac{1+\sqrt{5}}{2} = 1 + \sum_{n=0}^{\infty} (-1)^n \frac{1}{f_n f_{n+1}}.$$

We also have

$$\begin{aligned} \frac{1}{2}(\sqrt{5} + 1) &= [1, 1, 1, 1, \dots] =: [1, \overline{1}], \\ \sqrt{2} &= [1, 2, 2, 2, \dots] =: [1, \overline{2}], \\ \sqrt{5} &= [2, 4, 4, 4, \dots] =: [2, \overline{4}], \\ \sqrt{7} &= [2, 1, 1, 1, 4, 1, 1, 1, 4, \dots] =: [2, \overline{1, 1, 1, 4}]. \end{aligned}$$

Finally, as proved by Leonhard Euler (1707–1783),

$$\begin{aligned} e &= [2, 1, 2, 1, 1, 3, 1, 1, 4, 1, \dots] = [2, \overline{1, n, 1}]_{n=1}^{\infty}, \\ \frac{e^{k/2} + 1}{e^{k/2} - 1} &= [k, 3k, 5k, \dots] \quad k = 1, 2, \dots \end{aligned}$$

Instead, no formation rule for the continued fraction of π is known.

From Theorem 8.30 we readily infer that for the n -th convergents of x , we have

$$\left| x - \frac{p_n}{q_n} \right| \leq \frac{1}{q_n^2},$$

in particular, if α is irrational, then there exist infinitely many rationals p/q , p, q coprime, such that

$$\left| x - \frac{p}{q} \right| < \frac{1}{q^2}. \quad (8.30)$$

Moreover if x is rational, $x = a/b \neq p/q$, we have

$$\left| x - \frac{p}{q} \right| = \frac{|aq - bp|}{bq} > \frac{1}{bq}$$

thus, assuming (8.30), we get $q < b$ and in turn (8.30) has a finite number of solutions. In conclusion, we have the following.

8.33 Theorem (Dirichlet). α is irrational if and only if the inequality

$$|q\alpha - p| < \frac{1}{q}$$

is satisfied by infinitely many integers (p, q) , $q > 0$.

However, the estimate (8.30) is not sharp in the following sense. It can be easily proved that of any two consecutive convergents of x , one at least satisfies the inequality

$$\left| x - \frac{p}{q} \right| < \frac{1}{2q^2}. \quad (8.31)$$

In particular, there are infinitely many convergents which satisfy (8.31). More can be stated in this direction. For instance, it has been proved that for any three consecutive convergents to x one at least satisfies

$$\left| x - \frac{p}{q} \right| < \frac{1}{\sqrt{5}q^2}, \quad (8.32)$$

and therefore we have the following.

8.34 Theorem (Hurwitz). Every irrational x admits infinitely many rational approximations p/q such that

$$\left| x - \frac{p}{q} \right| < \frac{1}{\sqrt{5}q^2}.$$

The constant $\sqrt{5}$ in the Hurwitz theorem is sharp. It can be easily proved that if p_n/q_n denotes the irreducible representation of the n -th convergent of $(1 + \sqrt{5})/2$, then $q_n^2|x - p_n/q_n| \rightarrow 1/\sqrt{5}$, and that

$$\left| \frac{1 + \sqrt{5}}{2} - \frac{p_n}{q_n} \right| \leq \frac{1}{Aq_n^2}$$

holds only for a finite numbers of convergents if $A > \sqrt{5}$.

We conclude observing that (8.31), that is satisfied by at least half of the convergents, characterizes the convergents. We have the following.

8.35 Theorem. Let x be a positive real number. Assume that p, q are coprime integers. If (8.31) holds, then p/q is one of the convergents of the continued fraction development of x .

Proof. By assumption

$$\frac{p}{q} - x = \frac{\epsilon \alpha}{q^2}$$

where $\epsilon = \pm 1$ and $0 < \alpha < 1$. Let $[a_0, \dots, a_n]$ be the continued fraction expansion of p/q , n being such that $(-1)^{n-1} = \epsilon$, and let p_k/q_k be the k -th convergent of $[a_0, \dots, a_n]$, in particular $p = p_n$, $q = q_n$. Write x as

$$x = [a_0, \dots, a_n, z]$$

for a suitable z ; according to Proposition 8.19 we have

$$x = \frac{z p_n + p_{n-1}}{z q_n + q_{n-1}},$$

thus

$$\frac{\epsilon \alpha}{q_n^2} = \frac{p_n}{q_n} - x = \frac{z(p_n q_{n-1} - p_{n-1} q_n)}{z q_n + q_{n-1}}$$

i.e.,

$$\frac{q_n}{z q_n + q_{n-1}} = \alpha,$$

that is, $z \geq 1$. Thus the continued fraction expansion of z is a simple finite or infinite continued fraction $[b_0, \dots, b_n, \dots]$ with $b_0 \geq 1$. We then infer that

$$[a_0, \dots, a_n, b_0, \dots, b_n, \dots]$$

is a *simple* continued fraction, and

$$x = [a_0, \dots, a_n, z] = [a_0, \dots, a_n, b_0, \dots, b_n, \dots].$$

Hence $p/q = [a_0, \dots, a_n]$ is one of the convergents of the continued fraction development of x . \square

8.36 Periodic continued fractions. The reader may have already noticed that $\sqrt{2}$, $\sqrt{5}$, $\sqrt{7}$ have periodic expansions as continued fractions. This is a general fact. Furthermore, we have the following.

Theorem (Lagrange). *A periodic continued fraction is a quadratic surd, i.e., an irrational root of a quadratic equation with integral coefficients.*

e. Order of approximation and transcendental numbers

We say that α is *approximable by rationals to order n* if there is a constant $k = k(\alpha)$ depending on α for which

$$\left| \frac{p}{q} - \alpha \right| < \frac{k(\alpha)}{q^n}$$

has infinitely many solutions. As we have seen, *every rational is approximable to order 1* and not to a higher order, while *every irrational is approximable of order two* (see Theorem 8.34). This way we separate the irrationals in classes that are further and further away from rationals.

8.37 Theorem (Liouville). *A real algebraic number¹⁰ of degree n is not approximable to any order greater than n .*

¹⁰ We recall that an algebraic number is a solution of an algebraic equation with integral coefficients. If x satisfies an equation of degree n , but none of lower degree, then it is said to be of degree n .

Proof. Let ξ be a root of

$$f(\xi) := a_0\xi^n + a_1\xi^{n-1} + \cdots + a_n = 0$$

with $a_i \in \mathbb{Z}$, and let $p/q \neq \xi$ be an approximation of ξ . We can assume that p/q lies in $]\xi - 1, \xi + 1[$, and is nearer to ξ than any other root of $f(x) = 0$, so that $f(p/q) \neq 0$. Trivially there exists $M(\xi)$ such that

$$|f'(x)| \leq M(\xi) \quad \forall x \in [\xi - 1, \xi + 1],$$

and

$$\left| f\left(\frac{p}{q}\right) \right| = \frac{|a_0p^n + a_1p^{n-1}q + a_2p^{n-2}q^2 + \cdots|}{q^n} \geq \frac{1}{q^n},$$

since the numerator is a positive integer. Also by Lagrange's theorem

$$f(p/q) = f(p/q) - f(\xi) = \left(\frac{p}{q} - \xi\right)f'(x)$$

for some x lying between p/q and ξ . Therefore we conclude

$$\left| \frac{p}{q} - \xi \right| = \frac{|f(p/q)|}{|f'(x)|} \geq \frac{1}{M} \frac{1}{q^n}.$$

□

We can translate Liouville's theorem into the principle that rapidly converging sequences of rationals converge to a transcendental number, and simple irrationals like $\sqrt{5} - 1$ or $\sqrt{2}$ cannot be rapidly approximated by rationals: from the point of view of rational approximation the simplest numbers are the worst.

Liouville's theorem of course allows us to construct transcendental numbers easily.

8.38 Example. If

$$\xi = 0,110001000\dots = 10^{-1!} + 10^{-2!} + 10^{-3!} + \cdots = \sum_{k=1}^{\infty} 10^{-k!},$$

and

$$\xi_n = \sum_{k=1}^n 10^{-k!} =: \frac{p}{10^{n!}} =: \frac{p}{q},$$

we have

$$0 < \xi - \frac{p}{q} = \xi - \xi_n = \sum_{k=n+1}^{\infty} 10^{-k!} < 2 \cdot 10^{-(n+1)!} < 2q^{-N}$$

for $n > N$. Consequently ξ is not an algebraic number of degree less than N . N being arbitrary, we conclude that ξ is transcendental.

8.39 ¶¶. Show that the number $[1, 10, 10^2, 10^{3!}, 10^{4!}, \dots]$ is transcendental.

Of course, it is more difficult to decide whether a given number is transcendental or not. For instance, only in 1873 Charles Hermite (1822–1901) proved that π is transcendental, and in 1882 Carl von Lindemann (1852–1939) proved that e is transcendental, too. Even nowadays only few classes of transcendental numbers are known, for example,

$$e, \pi, \sin 1, \log 2, \log 3 / \log 2, e^\pi, 2^{\sqrt{2}}$$

are transcendental, but it is not known whether

$$2^e, 2^\pi, \pi^e$$

are transcendental or not; actually not even whether they are rational or irrational. We only state without proofs

8.40 Theorem (Roth, 1958). *The order of approximation of any algebraic rational is 2. In other words, given an algebraic number α and $k > 2$, there are only finitely many rational numbers p/q solving $|\alpha - p/q| < c/q^k$.*

8.41 Theorem (Lindemann–Weierstrass). *Let $\alpha_1, \dots, \alpha_n$ be n distinct complex numbers. If*

$$\beta_1 e^{\alpha_1} + \dots + \beta_n e^{\alpha_n} = 0, \quad \beta_j \in \mathbb{C},$$

then either at least one of the coefficients β_j is transcendental or all β_j vanish.

8.2 One-Dimensional Dynamical Systems

Roughly, a *dynamical system* consists of a set of possible states, together with rules that determine the present state in terms of the past states. The system is said to be *continuous* or *discrete* according to whether the rule is applied at continuous or discrete times.

Let I be a closed interval in \mathbb{R} or $I = \mathbb{R}$, and $f : I \rightarrow I$ be a mapping from I to I . A typical discrete dynamical system is

$$\begin{cases} x_0 = x, \\ x_{n+1} = f(x_n) \quad n \geq 0. \end{cases}$$

The sequence $f^n(x) := f \circ f \circ \dots \circ f(x)$ of the *iterates* of f is called the *orbit of x under f* . A point $x \in I$ such that $f(x) = x$ is called a *fixed point* of f . According to the principle of induction, the orbit $\{x_n\}$ is uniquely determined by the *initial value* x_0 ; nevertheless, for nonlinear

maps, even just quadratic maps, the sequence $\{x_n\}$ may have a behavior, or *dynamics*, quite complicated, with effects due to nonlinearity, such as absence of convergence, presence of oscillations, sensitive dependence on initial conditions which lead to a complex distribution of the values $\{x_n\}$, often referred to as *deterministic chaos*. Such dynamical systems occur in many contexts and, in particular, when *discretizing* differential equations or when studying *mathematical models*, as, for instance, population models.

8.2.1 Discretization and models

Given a smooth function $f : \mathbb{R} \rightarrow \mathbb{R}$, Cauchy's problem

$$\begin{cases} x(0) = x_0, \\ x'(t) = f(x(t)) \end{cases} \quad (8.33)$$

has a unique solution at least in a short interval $[0, T]$. Discrete *methods* allow us to approximate such a solution.

a. Euler's method

If we assume a small h as discrete approximation step, and replace in the differential equation the derivative with the differential quotient $(x(t+h) - x(t))/h$, we find

$$\frac{x(t+h) - x(t)}{h} = f(x(t)).$$

This suggests as approximate solution

$$x^{(h)}(t) := x_k + (ht - k)x_{k+1} \quad t \in [k/h, (k+1)/h), \quad k = 0, 1, 2, \dots$$

where

$$\begin{cases} x_0 = x_0, \\ x_{k+1} = x_k + hf(x_k). \end{cases} \quad (8.34)$$

In other words x_{k+1} is the arrival point after time h if we start from x_k and move with constant speed $f(x_k)$, while $x^{(h)}(t)$ is the linear interpolate. The discretization error

$$\epsilon(x(t), h) := \frac{x(t+h) - x(t)}{h} - f(x(t))$$

clearly is infinitesimal as $h \rightarrow 0$: in fact, we have the following.

8.42 Proposition. *The sequence of functions $\{x^{(h)}(t)\}$ converges uniformly to the solution $x(t)$ of (8.33) on every bounded interval in which $x(t)$ exists.*

This easily follows from

8.43 Proposition. Let $M := \sup_{]0,T[} |f(x)|$, $L := \sup_{]0,T[} |f'(x)|$ and $h := T/N$. Then

$$|x_N - x(T)| \leq \frac{M}{2}(e^{LT} - 1)h.$$

Proof. From the equation $x'(t) = f(x(t))$ we get

$$|x(t) - x(s)| \leq \int_s^t |f(x(\tau))| d\tau = M|s - t|, \quad (8.35)$$

$$x((k+1)h) = x(kh) + h \int_{kh}^{(k+1)h} f(x(s)) ds, \quad k = 0, 1, \dots, N-1.$$

If $x_0 := x(0)$, $x_{k+1} := x_k + hf(x_k)$ and $\delta_k := |x_k - x(kh)|$, we then infer

$$\begin{aligned} \delta_{k+1} &\leq \delta_k + h|f(x(kh)) - f(x_k)| + \int_{kh}^{(k+1)h} (f(x(s)) - f(x(kh))) ds \\ &\leq (1 + Lh)\delta_k + L \int_{kh}^{(k+1)h} |x(s) - x(kh)| ds \leq (1 + Lh)\delta_k + \frac{LM}{2}h^2. \end{aligned}$$

Iterating, we finally obtain

$$\delta_k \leq (1 + Lh)^k \delta_0 + \frac{LMh^2}{2} \sum_{j=0}^{k-1} (1 + Lh)^j = \frac{LMh^2}{2} \frac{(1 + Lh)^k - 1}{Lh},$$

since $\delta_0 = 0$, and for $k = N$ the conclusion. \square

Proof of Proposition 8.42. In fact for $s \in [kh, (k+1)h]$, we have

$$|x^{(h)}(s) - x(s)| \leq |x^{(h)}(s) - x_k| + |x_k - x(kh)| + |x(kh) - x(s)| \leq Ch,$$

since $|x^{(h)}(s) - x_k| \leq |x_{k+1} - x_k| \leq Mh$ by definition, $|x(kh) - x(s)| < Mh$ by (8.35) and $|x_k - x(kh)| \leq Ch$ by Proposition 8.43. \square

The first formal description of this numerical method for approximating solutions of an ODE seems due to Leonhard Euler (1707–1783), while the proof of convergence is due to Augustin-Louis Cauchy (1789–1857). However, it is worth noticing that the method had already been used by Sir Isaac Newton (1643–1727).

b. Runge-Kutta method

Differentiating the equation $x' = f(x)$ we find

$$\begin{aligned}x'(t) &= f(x(t)), \\x''(t) &= f'(x(t))f(x(t)), \\x'''(t) &= f''(x(t))f^2(x(t)) + (f'(x(t)))^2 f(x(t)),\end{aligned}$$

thus, defining

$$\Phi(x, h) := f(x) + \frac{h}{2}f'(x)f(x) + \frac{h^2}{6}(f''(x)f^2(x) + (f'(x))^2 f(x))$$

we may set up the iteration scheme

$$\begin{cases} x_0 = x_0, \\ x_{k+1} = x_k + h\Phi(x_k, h). \end{cases}$$

Using the second order Taylor polynomial, similarly to the above one can show that this scheme converges and yields a better approximation than Euler's method. And we can get even better approximations using higher order terms.

But, from the numerical point of view, the computation of high derivatives is costly since it corresponds to computing a function at many points with higher accuracy. Therefore, while in principle we can get better and better approximations using higher order Taylor polynomials, this is not convenient as it leads to algorithms which are not very efficient. Thus, let us come back to the question of a good choice of $\phi(x, h)$ in the iteration

$$x_{k+1} = x_k + h\Phi(x_k, h).$$

We have

$$\begin{aligned}x(t+h) &= x + hf(x) + \frac{h}{2}f'(x)f(x) + o(h^2), \\ \Phi(x, h) &= \Phi(x, 0) + h\Phi'(x, 0) + O(h^2),\end{aligned}$$

the discretization error is then

$$\epsilon(x, h) = f(x) + \frac{h}{2}f'(x)f(x) + o(h^2) - \Phi(x, 0) - h\Phi'(x, 0) + o(h^2),$$

that is, of order $o(h^2)$, if we have

$$f(x) + \frac{h}{2}f'(x)f(x) - \Phi(x, 0) - h\Phi'(x, 0) = 0. \quad (8.36)$$

For example the choice

$$\Phi(x, h) := Af(x) + Bf(x + Chf(x)), \quad \text{with } A + B = 1, \quad BC = \frac{1}{2}$$

gives a family of solutions of (8.36) which yield approximating methods of order two

- The *modified Euler method* or *middle point method* corresponds to

$$A = 0, \quad B = 1, \quad C = 1/2.$$

- The *method of Heun* corresponds to

$$A = 1/2, \quad B = 1/2, \quad C = 1.$$

Similarly, one can construct methods of approximation of order 3, 4 or higher. The *method of Runge–Kutta* is a fourth order method defined by

$$\Phi(x, h) := \frac{1}{6}(m_1 + 2m_2 + 2m_3 + m_4),$$

$$m_1 = f(x), \quad m_2 = f\left(x + \frac{h}{2}m_1\right),$$

$$m_3 = f\left(x + \frac{h}{2}m_2\right), \quad m_4 = f(x + hm_3).$$

8.44 ¶¶. The *global error of approximation* is defined by

$$E(h) = \sup_{[0, T]} |x(t) - x^{(h)}(t)|.$$

Show that in the case of Heun's method and Euler's method $E(h) = O(h^2)$, while in the case of Runge–Kutta $E(h) = O(h^4)$.

c. Models

Processes of the real world are often mathematically modelled in order to capture some of their characteristic and/or relevant aspects; from this point of view a model that is too close to reality may happen to be intractable and consequently useless, while a model which, though far from reality, identifies relevant specific aspects may be very useful: modelling is a kind of compromise. Industrial mathematics, biology, economics and social sciences are the context in which new models develop according to specific needs. Though fascinating, discussing even a few examples is not possible here, therefore we confine ourselves to presenting very briefly two classical examples in population dynamics: the *logistic model* and *Lotka–Volterra models*.

8.45 The logistic model. Let x_0 denote the initial size of a population and let $\{x_n\}$ be the size after n years. The rate of change is then

$$\frac{x_{n+1} - x_n}{x_n}.$$

If such a rate is constant, say α , the dynamics is described by

$$x_{n+1} = (1 + \alpha)x_n;$$

the size of the population after n years is $x_n = (1 + \alpha)^n x_0$, that is, the population increases exponentially if $\alpha > 0$. Such a model describes the

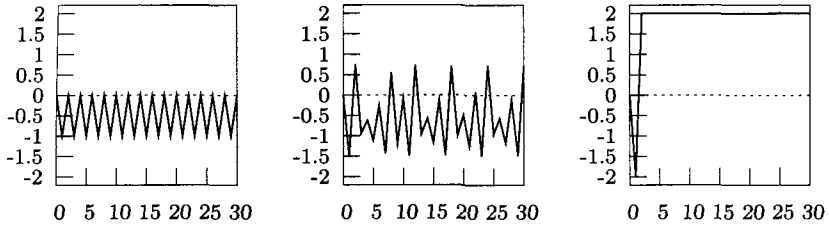


Figure 8.11. The iterates of $f(x) = x^2 + c$ starting from $x_0 = 0$ with, from the left $c = -1$, $c = -1.5$, $c = -2$.

ideal situation in which no external influence is present. More realistic models should take into account influences of the environment. In 1845 P. F. Verhulst, starting from the assumption that the environment may only allow the survival of a threshold population P_0 , that we take to be 1, formulated the hypothesis that the rate of change is proportional to $1 - x_n$. In this case the dynamics becomes

$$x_{n+1} = (1 + \alpha)x_n - \alpha x_n^2. \quad (8.37)$$

This is the *logistic model*; see, e.g., Section 4.2 of [GM1].

8.46 ¶. Show that Euler's method with step $h = 1$ for the equation

$$x' = \alpha(x - x^2)$$

leads to (8.37).

8.47 Lotka–Volterra models. These are models often used to simulate the interaction between two or more populations. In the case of two species, because of the finiteness of resources, the rate of change, per individual, is adversely affected by high levels of its own species (as in the logistic model) and by the other species with which it is in competition. We have

$$\frac{x'}{x} = A(E - x) - By$$

and, a similar equation for the second population y , i.e.,

$$\begin{cases} x' = Ax(E - x) - Bxy, \\ y' = Cy(F - y) - Dxy. \end{cases}$$

A special case is the so-called *predator-prey* model

$$\begin{cases} x' = ax - bxy, \\ y' = -cy + dxy \end{cases}$$

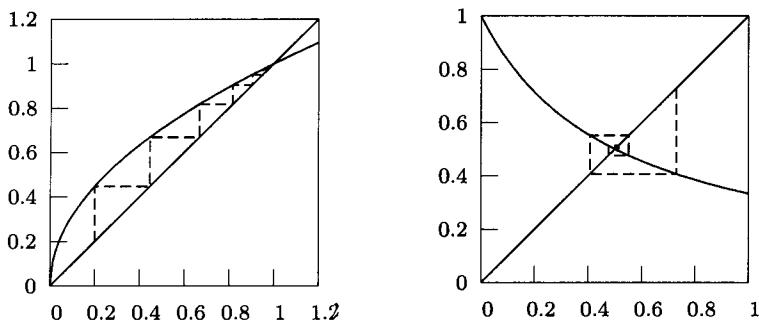


Figure 8.12. (a) The iterates of (a) \sqrt{x} starting from $x_0 = 0.2$ and of (b) $1/(2x + 1)$ starting from $x_0 = 0.2$.

that, in its discrete version, becomes

$$\begin{cases} x_{k+1} = (1 + a)x_k - bx_k y_k, \\ y_{k+1} = (1 - c)y_k + dx_k y_k. \end{cases}$$

8.2.2 Examples of one-dimensional dynamics

In this section we illustrate some typical features of one-dimensional discrete processes. Let

$$x_{n+1} = f(x_n),$$

be such a system, $f : \mathbb{R} \rightarrow \mathbb{R}$ being a given smooth function. The orbit of an initial point x_0 is given by the sequence $\{x_n\}$,

$$x_n = \underbrace{f \circ f \circ f \circ \cdots \circ f}_{n\text{-times}}(x_0), \quad \text{or simply} \quad x_n = f^n(x_0).$$

A graphic representation of the sequence $\{x_n\}$ may be given by the points (n, x_n) in the plane, possibly linearly interpolated, see Figure 8.11.

Alternatively, we plot in the (x, y) plane the graph of $y = f(x)$. We travel vertically from (x_0, x_0) till $(x_0, f(x_0))$ on the graph of f , then horizontally until we reach the diagonal, $(f(x_0), f(x_0)) =: (x_1, x_1)$, and finally, we reiterate starting from (x_1, x_1) , (x_2, x_2) , etc., see Figure 8.12.

Our first two examples concern very simple dynamics.

a. Expansive dynamics

Let $k > 1$ and let

$$x_{n+1} = kx_n.$$

This simple dynamics, that we have already encountered several times, has a closed form description:

$$x_n := k^n x_0.$$

The $f^n(x_0)$ separate exponentially in time. Also starting from slightly different data x_0, y_0 , then $x_n - y_n = k^n(x_0 - y_0)$, i.e., the orbits of x_0 and y_0 separate exponentially in time. Different of course is the case $0 < k < 1$: all initial points tend to zero: they are “attracted” by zero; zero is called a *sink*.

b. Contractive dynamics: fixed points

A *contraction map* or simply a *contraction* on an interval $I \subset \mathbb{R}$ is a map $f : I \rightarrow I$ which shrinks distances uniformly, i.e., for which there exists a constant L , $0 < L < 1$, such that $|f(x) - f(y)| \leq L|x - y| \forall x, y \in I$. Of course a contraction map is a Lipschitz map, in particular contraction maps are continuous on I .

8.48 Theorem (Contraction mapping theorem). *Let I be a closed subset of \mathbb{R} , for instance a closed interval, a closed half-line, a finite union of closed intervals, or \mathbb{R} , and let $f : I \rightarrow I$ be a contraction with contraction factor $L < 1$. Then f has a unique fixed point $x_0 \in I$. Moreover, the orbit of any point $x \in I$ converges at least exponentially to x_0 ,*

$$|f^n(x) - x_0| \leq \frac{L^n}{1-L} |f(x) - x| \quad \forall n. \quad (8.38)$$

Proof. Uniqueness. If $x, y \in I$ are two fixed points, we have

$$|x - y| = |f(x) - f(y)| \leq L|x - y|,$$

hence $x = y$, since $L < 1$.

Existence. For any $x \in I$, consider its orbit $\{x_n\}$, $x_n := f^n(x)$. We then have

$$|x_{k+1} - x_k| \leq L|x_k - x_{k-1}| \leq \cdots \leq L^k|x_1 - x_0| = L^k|f(x) - x|$$

hence, for $q \geq p \geq 1$,

$$|x_q - x_p| \leq \sum_{h=p}^{q-1} |x_{h+1} - x_h| \leq \sum_{h=p}^{q-1} L^h |f(x) - x| \leq |f(x) - x| \frac{L^p}{1-L}, \quad (8.39)$$

since $L < 1$. Since the right-hand side converges to zero as $p \rightarrow \infty$, $\{x_n\}$ is a Cauchy sequence and therefore converges to some $x_0 \in \mathbb{R}$. Passing to the limit in $x_{n+1} = f(x_n)$, we see that x_0 is a fixed point for f , thus proving that f has a fixed point, hence a unique fixed point, and that each orbit converges to it. Finally, (8.38) follows passing to the limit as $q \rightarrow \infty$ in (8.39). \square

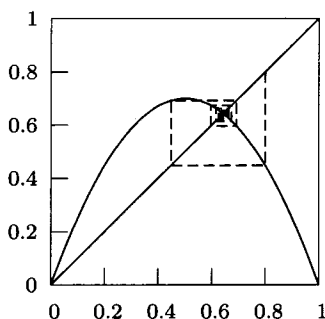


Figure 8.13. The iterates of $f(x) = 2.8x(1-x)$ starting from $x = 0.8$.

c. Sinks and sources

Let us begin by illustrating a number of phenomena associated to generic maps.

8.49 Example. Let $f(x) := 2x(1-x)$, $x \in [0, 1]$. Clearly, $f : [0, 1] \rightarrow [0, 1]$ has two fixed points, 0 and $1/2$, and the orbit of every point $x \in [0, 1]$, $x \neq 0$, converges to $1/2$; in fact, $\{f^n(x)\}$ is increasing, and passing to the limit in $x_{n+1} = f(x_n)$, necessarily, $f^n(x) \rightarrow 1/2$.

Trivially we have

8.50 Proposition. Let $f : I \rightarrow I$ be a continuous map, and let $\{f^k(x)\}$ be the orbit of $x \in I$. Then

- (i) if $f^k(x) \rightarrow p$ as $k \rightarrow +\infty$, then p is a fixed point, $f(p) = p$.
- (ii) if $f^n(x) = p$ for some n , and p is a fixed point, then $f^k(x) = p \ \forall k \geq n$.

Example 8.49 then suggests

8.51 Definition. Let $f : I \rightarrow I$ be a continuous map, and let p be a fixed point of f . We say that

- (i) p is stable if $\forall \epsilon > 0$ there is $\delta > 0$ such that $|f^n(x) - p| < \epsilon \ \forall n$ if $|x - p| < \delta$,
- (ii) p is a sink or an attracting point if there exists $\delta > 0$ such that $f^k(x) \rightarrow p$ for all x with $|x - p| < \delta$. If p is a sink, the basin of attraction of p is the subset of points on I whose orbits converge to p ,
- (iii) p is unstable if p is not stable, i.e., if for some $\epsilon_0 > 0$ there exists a sequence $\{x_k\}$ such that $x_k \rightarrow p$, and for each k an integer n_k such that $|f^{n_k}(x_k) - p| > \epsilon_0$,
- (iv) p is a source or a repelling point if for some $\epsilon_0 > 0$ and for each x such that $0 < |x - p| < \delta$ there is n_x such that $|f^{n_x}(x) - p| > \epsilon_0$.

If p is stable and a sink, then p is said to be an asymptotically stable fixed point.

Of course, by definition *sources are unstable fixed points*. Trivially the fixed point of a contraction on I is a stable fixed point and a sink, and its basin of attraction is the whole I .

In Example 8.49, 0 is a source, $1/2$ is a sink, and the basin of attraction of $1/2$ is $]0, 1]$. In general, we have

8.52 Theorem. Let p be a fixed point for a continuous map $f : I \rightarrow I$, and assume that f is differentiable at p .

- (i) If $|f'(p)| < 1$, then p is a stable fixed point and a sink, that is, p is asymptotically stable.
- (ii) If $|f'(p)| > 1$, then p is a source, in particular p is an unstable fixed point.

Proof. (i) Since $\lim_{x \rightarrow p} \frac{|f(x) - f(p)|}{|x - p|} = |f'(p)|$, there is $L < 1$ and $\delta > 0$ such that

$$|f(x) - p| = |f(x) - f(p)| \leq L|x - p| < |x - p|$$

if $x \in I$ and $|x - p| < \delta$. In particular, if $|\bar{x} - p| < \delta$, then $|f^n(x) - p| < \delta \forall n$, hence p is stable. Moreover

$$|f^{n+1}(x) - p| \leq L|f^n(x) - p| \quad \forall n,$$

and, by iteration, we therefore conclude that

$$|f^n(x) - p| \leq L^n|x - p|,$$

hence $\{f^n(x)\}$ converges exponentially to p as $n \rightarrow \infty$.

(ii) Similarly to (i), there is $L > 1$ and $\delta > 0$ such that $|f(x) - p| \geq L|x - p|$ if $|x - p| < \delta$. If p were not a source, for any $\epsilon > 0$ we could then find x arbitrarily close to p such that $|f^n(x) - p| < \epsilon \forall n$. Choosing $\epsilon = \delta$ we would find x such that $|f^n(x) - p| < \delta \forall n$ and setting $\sigma_n := |f^n(x) - p|$, $\sigma_n < \delta \forall n$. But on the other hand by assumption $\sigma_{n+1} \geq L\sigma_n \forall n$ i.e., $\sigma_n \rightarrow \infty$: a contradiction. \square

8.53 Remark. We notice that, if p is a fixed point of f , f is twice differentiable at p and $f'(p) = 0$, then the orbit $\{f^n(x)\}$ converges to p rapidly. In fact, in this case the second order Taylor formula yields $\sigma_{n+1} \leq M\sigma_n^2$, $\sigma_n := |f^n(x) - p|$, see (8.15).

d. Periodic orbits

In dependence on the parameter a , the dynamics of the logistic map $f_a(x) = ax(1 - x)$, $x \in [0, 1]$, is significantly different. Observe that the map $g(x) := ax(1 - x)$, $x \in \mathbb{R}$, has two fixed points in \mathbb{R} , 0 and $(a - 1)/a$.

For instance, if $0 < a < 1$, then $|f'_a(x)| \leq a < 1$ and f_a is a contraction with the unique fixed point $x = 0$. For $a = 1$ the map f_1 still has 0 as a unique fixed point, and it is easy to check that 0 is a sink. For $1 < a < 3$, f_a has two fixed points: 0 that is a source, and $(a - 1)/a$ that is a sink with $]0, 1]$ as basin of attraction.

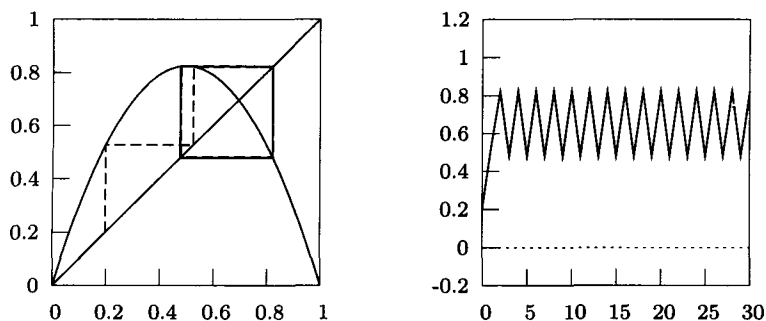


Figure 8.14. The iterates of $f(x) := 3.3x(1-x)$.

For $a = 3.3$, f_a has two fixed points, 0 and $22/33$, and both are sources. Therefore the orbits cannot converge. Numerical experiments show that the values of the iterates oscillate between two values of the order of 0.4794 and .8236, see Figure 8.14, and moreover $f(0.4794) = .8236$ and $f(.8236) = .4794$, modulus roundings. This suggests that the two alternating values are fixed points of f^2 . Actually, one easily proves that $g(x) := f \circ f(x)$ has exactly three fixed points $0, p_1, p_2$, Figure 8.15, with $p_1 := 0.4794\dots$ and $p_2 := .8236\dots$, and that both are sinks. The same holds if $3 < a < 1 + \sqrt{6} = 3.34494\dots$.

8.54 Definition. Let $f : I \rightarrow I$ be a continuous map. A k -periodic point is a fixed point of f^k , $k \geq 1$, that is not a fixed point of f^h for any $h < k$. A k -th periodic orbit is the orbit $\{f^n(p)\}$ of a k -periodic point p .

Of course, a k -th periodic orbit consists of k distinct points that are k -periodic points.

8.55 Example. The origin is the unique fixed point of $f(x) = -x$, $x \in [-1, 1]$; any other point is a 2-periodic point, since $f \circ f(x) = -(-x) = x$, and the 2-periodic orbit of x is the sequence $\{(-1)^n x\}$.

The *stability* of k -periodic orbits can now be discussed exactly in the same terms in which we discussed the stability of fixed points.

8.56 Definition. Let $f : I \rightarrow I$ be a continuous map on a closed interval I and let p be a k -periodic point of f . We say that the k -periodic orbit of p is a k -periodic sink or a k -periodic attractor if p is a sink for the k -th iterate f^k of f .

We say that the k -periodic orbit of p is a periodic source if p is a source for f^k .

Finally, we say that the k -periodic orbit of a k -periodic point p is stable, asymptotically stable, unstable if respectively p is a stable, asymptotically stable, unstable fixed point of f^k .

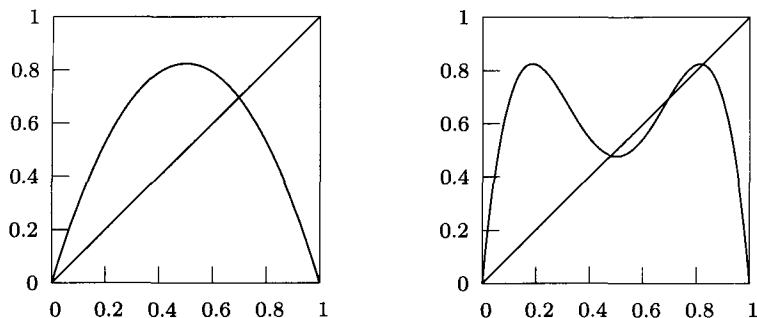


Figure 8.15. On the left $f(x) = 3.3x(1-x)$ and on the right $f^2(x)$.

8.57 ¶. Let $f : I \rightarrow I$. An orbit $\{f^n(x)\}$ is said to be *asymptotically k -periodic* if $|f^n(x) - f^n(p)| \rightarrow 0$ as $n \rightarrow \infty$ for some k -periodic point p . Show that p is a k -periodic sink if and only if there is $\delta > 0$ such that $\{f^n(x)\}$ is asymptotically periodic for all x with $|x - p| < \delta$.

Let p be a k -periodic point of f that we assume to be differentiable at p . Denote by $p_0, \dots, p_{k-1}, p_0 := p$, the k distinct values of the orbit $f^k(p)$; since we have

$$\begin{aligned} D(f^k)(p) &= f'(f^{k-1}(p))f'(f^{k-2}(p)) \cdots f'(f(p))f'(p) \\ &= f'(p_{k-1})f'(p_{k-2}) \cdots f'(p_0), \end{aligned}$$

from Theorem 8.48 we infer at once the following.

8.58 Theorem. Let $f : I \rightarrow I$ be differentiable. If $p_0, \dots, p_{k-1}, p_0 := p$, are the k values of a k -periodic orbit then

- (i) if $|f'(p_0)f'(p_1) \cdots f'(p_{k-1})| < 1$, the orbit is asymptotically stable, that is, the orbit is a sink and is stable.
- (ii) if $|f'(p_0)f'(p_1) \cdots f'(p_{k-1})| > 1$, the orbit of p is a source, hence unstable.

8.59 Example. In the case $f(x) = 3.3x(1-x)$, $x \in [0, 1]$, none of the two fixed points 0.4794 and 0.8236 is a fixed point of f . The corresponding 2-orbit is a periodic sink, since $|f'(0.4794)f'(0.8236)| < 1$, see Figures 8.14 and 8.15.

e. Periodic-doubling cascade transition to chaos

If we increase the parameter a in the logistic map, the dynamics becomes more and more complex. The so-called *bifurcation diagram* of f_a is plotted in Figure 8.16. The diagram has been obtained printing the parameter a as abscissa, computing the iterates $f_a^k(x_0)$ and plotting on the line $x = a$ the values of the k -orbit of x_0 for $200 < k < 400$. The values of x_0 and of $f_a^k(x_0)$ for $1 < k \leq 200$ have been neglected in order to eliminate the

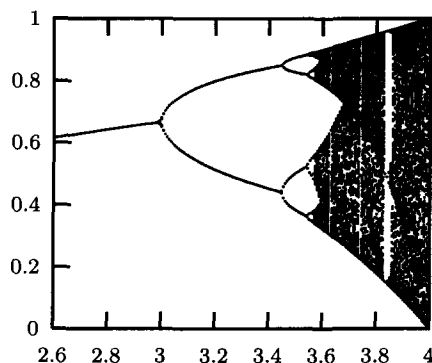


Figure 8.16. The iterates of the logistic map $f(x) := \lambda x(1-x)$.

dependence from the initial data as much as possible. The figure suggests that for $a = 3.45$, there is a 4-orbit which sinks: this can in fact be proved along the same lines as Theorem 8.58. As a increases, one gets an 8-orbit, a 16-orbit and actually an entire sequence of 2^n -orbits, $n = 1, 2, \dots$ which sink, until a reaches a limiting value $a_\infty := 3.5699456\dots$. Such a sequence is called a *periodic-doubling cascade* and is one of the *routes to chaos* since in fact, for $a > a_\infty$ the orbits appear to randomly fill out the entire interval or a subinterval: they are quite complicated, hard to describe and quite irregular. But before that *chaos*, there is some regularity, which in fact is *universal*. In fact, set $f_a := af(x)$ for any continuous map $f : [0, 1] \rightarrow [0, 1]$ with a unique maximum point, and $f(0) = f(1) = 0$, and denote by a_n the value at which the n -th bifurcation occurs. Then it has been proved that there is a constant δ , called the *Feigenbaum constant*, such that

$$\lim_{n \rightarrow \infty} \frac{a_n - a_{n-1}}{a_{n+1} - a_n} = \delta = 4.66920161\dots$$

This is an example of *order in the transition to chaos*.

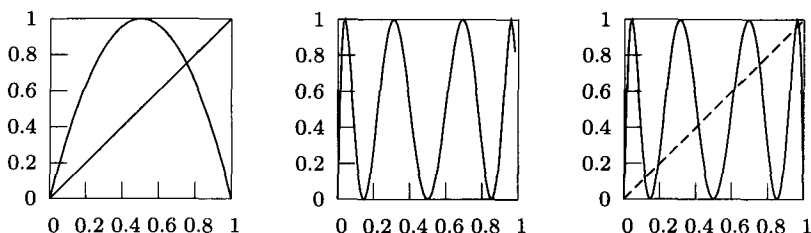


Figure 8.17. The map $f(x) := 4x(1-x)$ and the graphs of $f(x)$, $f^2(x)$, $f^3(x)$.

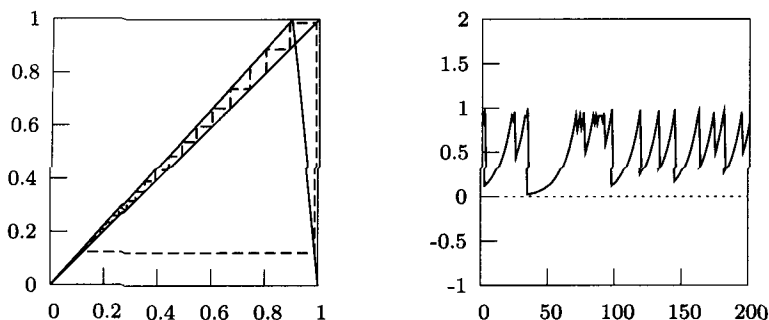


Figure 8.18. The intermittency phenomenon due to a thin channel.

8.60 Example. The map $f(x) = 4x(1-x)$, $x \in [0, 1]$, has two fixed points, $x = 0, 3/4$, but no sink. The map f_4^2 has four fixed points $x = 0, p_1, 3/4, p_2$, two of which are the fixed points of f and the other two are 2-periodic sinks $f(p_2) = p_1$ and $f(p_1) = p_2$. The map f_4^3 has eight fixed points: $0, 3/4$, the 2-periodic sinks are not fixed points for f_4^3 , the remaining six points form 2-orbits of period 3, see Figure 8.17. For more complicated orbits, compare Proposition 8.75.

f. The intermittency phenomenon

Consider the map

$$f(x) = \begin{cases} \alpha x & \text{if } x \in [0, 1/\alpha], \\ \frac{\alpha}{\alpha-1}(1-x) & \text{if } x \in]1/\alpha, 1]. \end{cases}$$

For $\alpha > 1$, one sees that f has no stable fixed point or orbit, and the orbits become quite complex; for a long period the process is regular, then the iterates oscillate around the unstable fixed point $x^0 := \alpha/(2\alpha-1)$, for some time after which they go away and the process restarts again more or less similarly, see Figure 8.19.

A similar phenomenon can be observed for the maps $g_\alpha(x) := g(x-\alpha)$ in Figures 8.19 and 8.20.

For $\alpha > \alpha_0$, x_n quickly get close to a fixed point, while for $\alpha < \alpha_0$ and $\alpha \simeq \alpha_0$, the iterates will remain for some time in a “channel,” then they jump outside the channel, after which they move back in for a long interval which in general depends on the point at which the iterate enters the channel.

In the formulas of logistic maps $f_\alpha = \alpha x(1-x)$ and of the map above, when α varies, as we have seen we experience a transition from a regular regime to a “chaotic” regime: they may be regarded as two examples of *transition to chaos*.

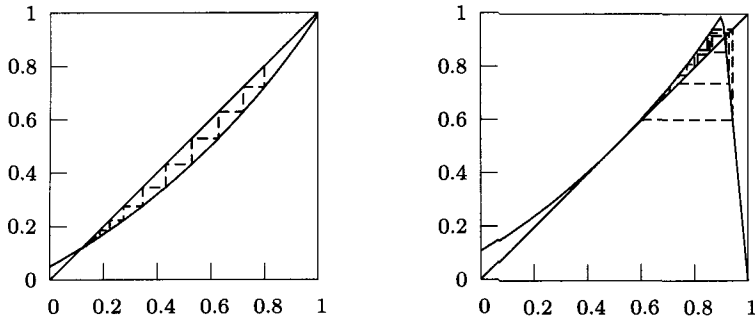


Figure 8.19. The iterates of (a) $g(x - 0.6)$ and of (b) $g(x - 0.5)$ where $g(x) = e^x - 1/2$ until it reaches 1 and then decays linearly.

g. Ergodic dynamics

Let us consider the dynamical system

$$x_{n+1} = x_n + \omega \pmod{1} \quad (8.40)$$

associated to the map $\varphi_\omega : [0, 1] \rightarrow [0, 1]$, $\varphi_\omega(x) = x + \omega \pmod{1}$, that maps x into the fractional part of $x + \omega$.

Clearly, it can be regarded as a dynamical system on the circle S^1 . In fact, if we identify \mathbb{R}/\mathbb{Z} with S^1 with the map $t \rightarrow \exp(i2\pi t)$ and set $z_n := e^{i2\pi x_n}$, we have

$$z_n = e^{i2\pi n\omega} z, \quad z := e^{i2\pi x}.$$

In other words, z_{n+1} is the point z_n rotated counterclockwise to the angle $2\pi\omega$.

If ω is *rational*, $\omega = p/q$, p, q coprime, the orbits of the system are all q -periodic and consist of a finite number of points $x_j := x + \frac{j}{q} \pmod{1}$.

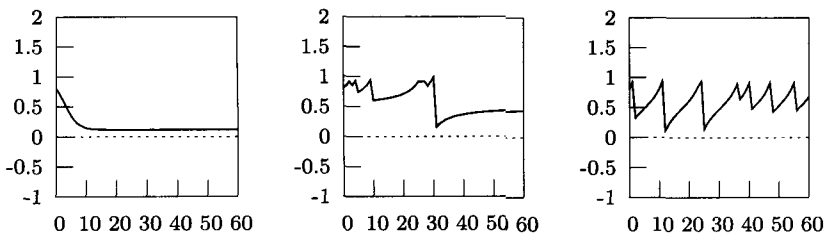


Figure 8.20. Here $g(x) = e^x - 1/2$ until it reaches 1 and then decays linearly. From the left, the iterates of $g(x - 0.6)$, $g(x - 0.5)$, and of $g(x - 0.45)$.

8.61 Theorem (Jacobi). *If ω is irrational, all orbits of φ_ω are dense in $[0, 1]$.*

Proof. Let ω be irrational and let $x \in [0, 1]$. First we notice that the points $\{\varphi_\omega^n(x)\}$ are distinct. In fact, if $\varphi_\omega^n(x) = \varphi_\omega^m(x)$, then $(n - m)\omega \in \mathbb{Z}$, consequently $n = m$, ω being irrational. The infinite distinct points of an orbit admit a convergent subsequence, by the Bolzano–Weierstrass theorem, hence for any $\epsilon > 0$, we can find two distinct elements $\varphi_\omega^n(x)$ and $\varphi_\omega^m(x)$ such that $|\varphi_\omega^n(x) - \varphi_\omega^m(x)| < \epsilon$. Since φ_ω preserves the distances on the circle, we deduce

$$|\varphi_\omega^p(x) - \varphi_\omega(x)| < \epsilon \quad \text{for } p = |n - m|,$$

consequently

$$0 < |\varphi_\omega^p(x) - \varphi_\omega(x)| = |\varphi_\omega^{(k+1)p}(x) - \varphi_\omega^{kp}(x)| < \epsilon.$$

We conclude that the sequence $\varphi_\omega^p(x), \varphi_\omega^{2p}(x), \varphi_\omega^{3p}(x), \dots$ divides S^1 in arcs of length uniformly bounded from below and not larger than ϵ : this shows that the orbit of x is dense. \square

The dynamics of (8.40), though quite complex, has relevant regularity properties. The *time mean* of a continuous function $f : [0, 1] \rightarrow \mathbb{R}$ or $f : [0, 1] \rightarrow \mathbb{C}$, also called an *observable* along the orbit, is defined as the limit

$$f^*(x) := \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N f(\varphi^n x)$$

if it exists, while its mean is called the *phase mean*

$$\bar{f} := \int_0^1 f(x) dx.$$

8.62 Theorem (Ergodic theorem). *Let ω be irrational. Then the limit $\varphi_\omega^*(x)$ exists for all x and $\varphi_\omega^*(x) = \bar{\varphi}_\omega$.¹¹*

In other words, the time mean of an observable along the orbit is independent of the orbit itself, equivalently on the initial value, and equals the phase mean.

Theorem 8.62 can be obtained as a consequence of the Hermann Weyl (1885–1955) theorem on the uniform equidistribution of the fractional parts of $\omega, 2\omega, 3\omega, \dots, n\omega$ for n large, if ω is irrational.

8.63 Theorem (Weyl). *Let ω be an irrational.*

¹¹ Actually one refers to a dynamics associated with a map $f : [0, 1] \rightarrow [0, 1]$ as to an *ergodic dynamics* if $f^*(x) = \bar{f}$ for almost every point $x \in [0, 1]$ in the sense of Lebesgue.

(i) For every continuous and 1-periodic function, we have

$$\frac{1}{N} \sum_{n=1}^N f(n\omega) \rightarrow \int_0^1 f(t) dt. \quad (8.41)$$

(ii) If $0 \leq a \leq b \leq 1$, then

$$\frac{\#\{n \mid 1 \leq n \leq N, n\omega \in [a, b]\}}{N} \rightarrow b - a$$

where, we recall, $\#A$ denotes the number of elements of A .

The proof of the Weyl theorem we present here uses the following density result that we do not prove.

8.64 Theorem. *Trigonometrical polynomials in $[0, 1]$ are dense in the class of continuous and 1-periodic functions, i.e., given a continuous function $f : [0, 1] \rightarrow \mathbb{R}$ with $f(0) = f(1)$, and a positive $\epsilon > 0$, there is a 1-periodic trigonometric polynomial such that*

$$|f(t) - P(t)| < \epsilon \quad \forall t \in [0, 1].$$

Proof of Theorem 8.63. (i) For increasingly complex f we shall prove that

$$G_N(f) := \frac{1}{N} \sum_{n=1}^N f(n\omega) - \int_0^1 f(t) dt \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

(a) If $f(t) = 1$, then clearly $G_N(1) = \frac{1}{N} \sum_{n=1}^N 1 - \int_0^1 1 dt = 0$.

(b) Suppose $f(t) = \exp(i2\pi kt)$, $k \in \mathbb{Z}$, $k \neq 0$, so that $\int_0^1 f(t) dt = 0$. Since ω is irrational, $\exp(i2\pi k\omega) \neq 1$, hence

$$\begin{aligned} |G_N(f)| &= \frac{1}{N} \left| \sum_{n=1}^N e^{i2\pi n k \omega} \right| = \frac{1}{N} \left| e^{i2\pi k \omega} \sum_{n=0}^{N-1} e^{i2\pi n k \omega} \right| \\ &= \frac{1}{N} \left| e^{i2\pi k \omega} \frac{1 - e^{i2\pi N k \omega}}{1 - e^{i2\pi k \omega}} \right| \leq \frac{1}{N} \frac{2}{|1 - e^{i2\pi k \omega}|} \rightarrow 0 \text{ as } N \rightarrow \infty. \end{aligned}$$

(c) Suppose now that f is a trigonometric polynomial of period 1, $P(t) = \sum_{k=-p}^p c_k e^{i2\pi kt}$. We have $G_N(P) = \sum c_k G_N(\exp(i2\pi kt))$ hence (ii) yields $G_N(P) \rightarrow 0$ as $N \rightarrow \infty$.

(d) For a continuous and 1-periodic function $f : \mathbb{R} \rightarrow \mathbb{C}$, given $\epsilon > 0$, by the density theorem, Theorem 8.64, we find a trigonometric polynomial $P(t)$ such that $|f(t) - P(t)| < \epsilon \forall t$. Thus

$$|G_N(f) - G_N(P)| \leq 2\epsilon \quad \forall N, \forall t.$$

According to (c), we can find N_0 such that $|G_N(P)| < \epsilon$ for all $N \geq N_0$. Therefore, we conclude for $N \geq N_0$,

$$|G_N(f)| \leq |G_N(P)| + |G_N(f) - G_N(P)| \leq 3\epsilon,$$

i.e., $G_N(f) \rightarrow 0$ as $N \rightarrow \infty$.

(ii) Given $\epsilon > 0$, let f_- and f_+ be two continuous functions such that



Figure 8.21. Aleksandr Lyapunov (1857–1918).

$$\begin{aligned}
 f_-(t) &\leq 1 \leq f_+(t) && \forall t \in [a, b], \\
 f_-(t) &= 0, && f_+(t) \geq 0 \quad \forall t \notin [a, b], \\
 (b-a) - \epsilon &\leq \int_0^1 f_- dt \leq \int_0^1 f_+ dt \leq (b-a) + \epsilon.
 \end{aligned}$$

Trivially

$$\sum_1^N f_-(n\omega) \leq \#\{n \mid 1 \leq n \leq N, n\omega \in [a, b]\} \leq \sum_1^N f_+(n\omega).$$

On the other hand, by (i), for $N \geq N_0 = N_0(\epsilon)$ we have $|G_N(f_+)|, |G_N(f_-)| \leq \epsilon$, hence

$$\int_0^1 f_- dt - \epsilon \leq \frac{\#\{n \mid 1 \leq n \leq N, n\omega \in [a, b]\}}{N} \leq \int_0^1 f_+ dt + \epsilon$$

and therefore

$$(b-a) - 2\epsilon \leq \frac{\#\{n \mid 1 \leq n \leq N, n\omega \in [a, b]\}}{N} \leq (b-a) + 2\epsilon.$$

□

We notice that the conclusions of Theorem 8.62 would be false if ω were rational. Therefore Theorem 8.62 characterizes the irrationals as the reals ω for which either (8.41) holds or the fractional parts of $\{n\omega\}$ are equidistributed.

8.2.3 Chaotic dynamics

Let us discuss now some of the characteristic features of chaos.

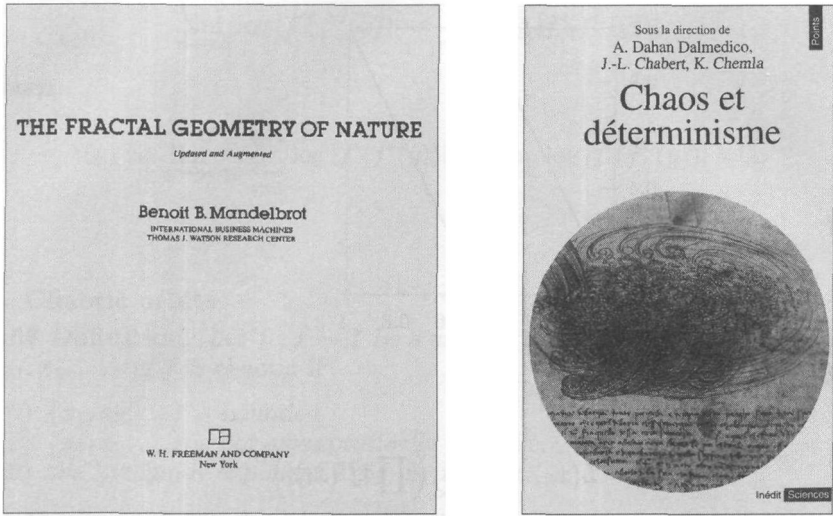


Figure 8.22. Frontispieces of two stimulating books.

a. Sensitive dependence on initial conditions and the Lyapunov exponent

8.65 Definition. Let $f : I \rightarrow I$ be a map defined on a closed interval $I \subset \mathbb{R}$. We say that a point $x_0 \in I$ has sensitive dependence on initial conditions or is a sensitive point if there exist $\epsilon_0 > 0$ and sequences $\{x_k\}$ and $\{n_k\}$ such that $x_k \rightarrow x_0$ and $|f^{n_k}(x_k) - f^{n_k}(x_0)| \geq \epsilon_0$.

It is not difficult to show that sources of any power of f are sensitive points.

In the case in which points near x_0 become separated by the action of the map f , a measure of such a separation is provided by the *Lyapunov exponent*. If the separation is exactly exponential, that is $|f^n(x) - f^n(x_0)| = q^n|x - x_0|$, the *Lyapunov number* is q . In the general case, the *Lyapunov number* at x_0 is defined by

$$L(x_0) := \lim_{n \rightarrow \infty} \limsup_{x \rightarrow x_0} \left(\frac{|f^n(x) - f^n(x_0)|}{|x - x_0|} \right)^{1/n}$$

and the *Lyapunov exponent*,

$$\lambda(x_0) := \log L(x_0),$$

is a measure of the (exponential) separation of the orbits starting close to x_0 , provided, of course, the limit as $n \rightarrow \infty$ exists.

If f is differentiable near x_0 , the Lagrange mean value theorem yields

$$L(x_0) := \lim_{n \rightarrow \infty} \left| (f^n)'(x_0) \right|^{1/n}$$

or, since $(f^n)'(x_0) = \prod_{i=1}^n f'(x_i)$, $x_i := f^i(x_0)$,

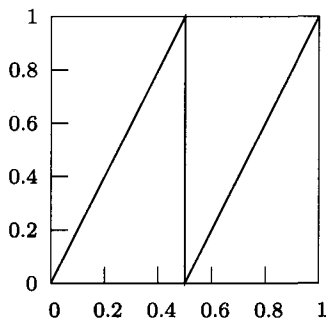


Figure 8.23. Bernoulli's shift.

$$L(x_0) := \lim_{n \rightarrow \infty} \left(\prod_{i=1}^n |f'(x_i)| \right)^{1/n}$$

and consequently

$$\lambda(x_0) = \log L(x_0) := \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log |f'(x_i)|.$$

In particular, the Lyapunov exponent $\lambda(x_1)$ of a fixed point x_1 of a smooth function f is $\log |f'(x_1)|$, while the Lyapunov exponent of a k -periodic orbit, at each of the values of the orbit, $x_1, x_2, \dots, x_k, x_{i+1} = f(x_i), x_i \neq x_j \forall i \neq j, x_k = x_1$, is

$$\lambda(x_1) = \lambda(x_2) = \dots = \lambda(x_k) = \frac{1}{k} \sum_{i=1}^k \log |f'(x_i)|.$$

8.66 Proposition. Let $f : I \rightarrow I$ be of class C^1 , and let $\{x_n\}, \{y_n\}, x_n = f^n(x), y_n = f^n(y)$ be the orbits of x and y respectively. Suppose that

- (i) $\{x_n\}$ and $\{y_n\}$ are asymptotic to each other, $|x_n - y_n| \rightarrow 0$ as $n \rightarrow \infty$,
- (ii) $f'(x_n) \neq 0, f'(y_n) \neq 0 \forall n$,
- (iii) $|f'(x_n)| \rightarrow \lambda \in \mathbb{R}$.

Then the Lyapunov exponents of f at x and y exist and $\lambda(x) = \lambda(y) = \lambda$.

Proof. In fact, Cesàro's theorem (see Example 2.56) yields

$$\lambda(x) := \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log |f'(f^i(x))| = \lim_{n \rightarrow \infty} \log |f'(f^n(x))| = \lambda.$$

Since the two orbits are asymptotic,

$$\lim_{n \rightarrow \infty} \log |f'(f^n(x))| = \lim_{n \rightarrow \infty} \log |f'(f^n(y))|$$

hence

$$\lambda(y) := \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log |f'(f^i(y))| = \lim_{n \rightarrow \infty} \log |f'(f^n(y))| = \lambda.$$

□

b. Chaotic orbits

8.67 Definition. Let $f : I \rightarrow I$ be a smooth map. We say that the orbit $\{x_1, x_2, \dots\}$ of f is chaotic if

- (i) $\{x_1, x_2, \dots\}$ is bounded,
- (ii) $\{x_1, x_2, \dots\}$ is not asymptotically periodic,
- (iii) the Lyapunov exponent $\lambda(x_1)$ is positive.

More complex and with less unanimous agreement is the definition of *chaotic dynamical system*. Usually, the presence of bounded orbits is required, with exponential separation and density to grant irreducibility, that is, that we are not in the presence of two assembled independent dynamical systems. We now illustrate a few examples.

c. Bernoulli's shift

Let us consider the process $x_{n+1} = \sigma(x_n)$ associated to the map

$$\sigma(x) = 2x \bmod (1), \quad x \in [0, 1],$$

see Figure 8.23. In order to describe its action, it is convenient to work with numbers in $[0, 1]$ in their binary representation. We write

$$x = \sum_{i=1}^{\infty} a_i 2^{-i} = 0.a_1 a_2 a_3 \dots$$

where a_i has the value 0 or 1. For $x < 1/2$, we have $a_1 = 0$ while $x \geq 1/2$ implies $a_1 = 1$. Therefore

$$\sigma(x) = \begin{cases} 2x & \text{if } a_1 = 0, \\ 2x - 1 & \text{if } a_1 = 1, \end{cases}$$

or

$$\sigma(0.a_0 a_1 a_2 \dots) = 0.a_1 a_2 a_3 \dots,$$

that is, the action of σ on the binary representation of x is to delete the first digit and shift the remaining sequence to the left.

The process σ shows sensitive dependence on the initial condition. If two numbers differ starting only from the n -th digit, such a difference becomes amplified under the action of σ^n by 2^n : the n -th iterates differ

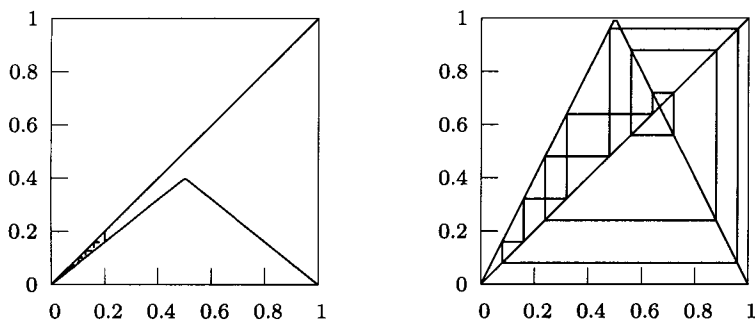


Figure 8.24. The triangular map $f_1(x)$ and its third iterate $f_1^3(x)$.

in the first digit. More precisely, we compute the Lyapunov exponent at a point x whose orbit never goes through $0, 1/2, 1$ as

$$\lambda(x) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log |f'(x_i)| = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log 2 = \log 2.$$

On the other hand periodic points are the numbers with a repeating binary representation, and the asymptotically periodic orbits are those with an initial point with a repeating binary representation starting from a suitable digit. Therefore an asymptotically periodic orbit starts at x if and only if x is rational. Hence the orbits starting from an irrational are bounded, are not asymptotically periodic and have Lyapunov exponent $\log 2 > 0$. So we conclude

8.68 Theorem. $x \in [0, 1]$ has a chaotic orbit under Bernoulli's shift if and only if x is irrational.

One can also prove the following properties of Bernoulli's shift σ :

- (i) The sequence of the iterates $\sigma^n(x_1)$ has the same random properties as the successive tosses of a coin. In fact, a sequence of coin tossings is equivalent to the choice of a point in $[0, 1]$,
- (ii) One can show¹² that almost all¹³ irrationals in $[0, 1]$ contain in their binary representation any finite sequence of digits infinitely often and uniformly distributed, i.e.,

$$\frac{1}{N} \sum_{n=1}^N F(\sigma(nx)) \rightarrow \int_0^1 F(t) dt.$$

Bernoulli's shift contrasts with the additive process we discussed before,

$$x_{n+1} = x_n + \omega \pmod{1}, \quad \omega \text{ irrational}, \quad (8.42)$$

¹² See, e.g., G.H. Hardy, E.M. Wright *The theory of numbers* Oxford University Press, Oxford 1938.

¹³ in the sense of Lebesgue.

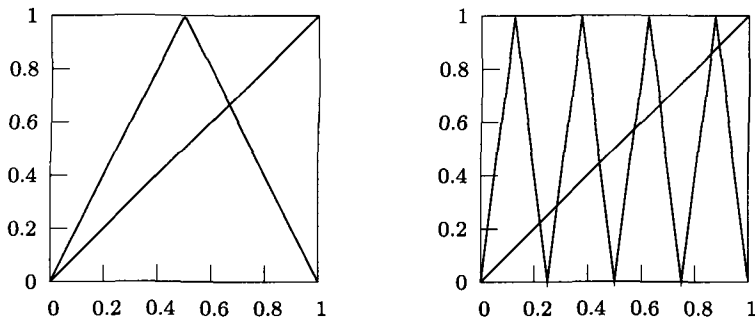


Figure 8.25. The triangular map $f_1(x)$ and its third iterate $f_1^3(x)$.

that we cannot qualify as chaotic. In fact, it has no periodic orbits and consequently no asymptotically periodic orbit, each orbit is dense, but the Lyapunov exponent of each orbit is zero.

8.69 Definition. A bounded orbit that is not asymptotically periodic and does not show sensitive dependence on the initial data is called almost-periodic.

The orbits of (8.42) are almost periodic.

d. The triangular map

Consider the family of triangular maps

$$f_r(x) := \begin{cases} 2rx & \text{if } x < 1/2 \\ 2r(1-x) & \text{if } x \geq 1/2 \end{cases}, \quad x \in [0, 1].$$

For $r < 1/2$, $x^* = 0$ is the only stable fixed point to which all points in $[0, 1]$ are attracted. For $r > 1/2$ two unstable fixed points (sources) emerge, and the behaviors of the process for $1/2 \leq r \leq 1$ and $r = 1$ are similar. The map f_1 shows sensitive dependence on initial conditions. In fact, the n -th iterate of f_1 is piecewise linear, with slopes $\pm 2^n$ except at the points $j \cdot 2^{-n}$, $j = 0, 1, \dots, 2^n$. Consequently the separation of “almost all points” x_0 grows exponentially with Lyapunov exponent $\lambda(x_0) = \log 2$. In general, the Lyapunov exponent of f_r is for almost all points x_0 ,

$$\lambda(x_0; f_r) = \log 2r,$$

and, for $r > 1/2$, we have $\lambda(x_0, f_r) > 0$, that is we “lose information” on the position of x_0 after n iterations, while for $r < 1/2$, we have $\lambda(x_0, f_r) < 0$ and we “gain information” as the iterates of x_0 converge to 0.

On the other hand, one can show that the periodic orbits do not attract any orbit, inferring this way the following.

8.70 Proposition. *The triangular map f_1 has infinitely many chaotic orbits.*

e. Conjugate maps

8.71 Definition. *Two continuous maps f and g are said to be conjugate if there is a continuous and 1-to-1 continuous change of variable C such that $C \circ f = g \circ C$.*

8.72 Proposition. *Let f and g be conjugate by C . If x is k -periodic for f , then $C(x)$ is k -periodic for g . If moreover f , g and C are of class C^1 and never vanish along the k -periodic orbit of x , then*

$$(g^k)'(C(x)) = (f^k)'(x).$$

8.73 Proposition. *The triangular map f_1 and the logistic map $g(x) = 4x(1-x)$ are conjugate.*

8.74 ¶. Show Proposition 8.73. [Hint: For $x \in [0, 1/2]$ choose $C(x) := \frac{1-\cos \pi x}{2}$.]

Taking into account Propositions 8.72 and 8.73 one could also show the following.

8.75 Proposition. *Let $g(x) := 4x(1-x)$ be the logistic map.*

- (i) *All periodic points of g are sources.*
- (ii) *g has chaotic orbits.*

8.76 ¶¶. Show that the process associated to the logistic map g is ergodic and for all F

$$\frac{1}{N} \sum_1^N F(g^n x) \rightarrow \int_0^1 \frac{F(t)}{\pi \sqrt{t-t^2}} dt.$$

8.77 ¶. Show that

- (i) the maps $(\alpha+1)x - \alpha x^2$ and $x^2 + c$, $c := \frac{1-\alpha^2}{2}$, are conjugate,
- (ii) the maps $\alpha x(1-x)$ and $x^2 + c$, $c = \frac{\alpha}{2}(1 - \frac{\alpha}{2})$, are conjugate.

[Hint: (i) $\varphi(x) := \frac{1+\alpha}{2} - \alpha x$, (ii) $\varphi(x) := \frac{\alpha}{2} - \alpha x$.]

8.2.4 Chaotic attractors, basins of attraction

Let us start with a few definitions.

The *forward limit* of x is the set of points the orbit converges to, that is, the set of points to which the orbit with initial condition x comes infinitely often arbitrarily close to, i.e.,

$$\omega(x) := \left\{ y \mid \liminf_{n \rightarrow \infty} |f^n(x) - y| = 0 \right\}.$$

The following is trivial:

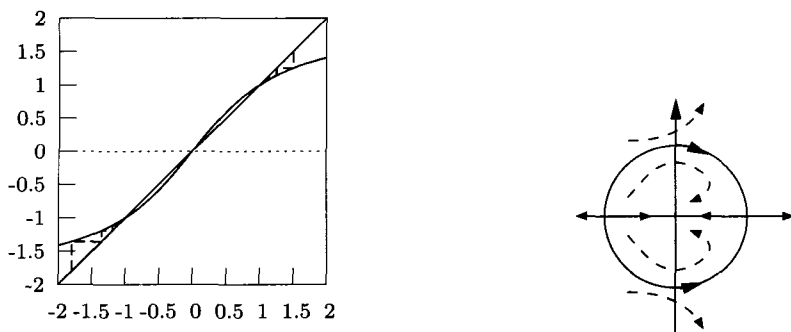


Figure 8.26. (a) The map $f(x) = \frac{4}{\pi} \arctan x$. (b) The map in Example 8.80.

- (i) if $\{f^n(x)\}$ converges to p , then $\omega(x) = \{p\}$,
- (ii) if x is a fixed point, $\omega(x) = \{x\}$, actually, if $\omega(x)$ contains a stable fixed point \bar{x} , then $\omega(x) = \{\bar{x}\}$,
- (iii) if $y \in \omega(x)$, then $f^n(y) \in \omega(x) \forall n \geq 0$,
- (iv) if x is a k -periodic point, then $\omega(x)$ is the set of values of the periodic orbit originating from x .

When the forward limit is a set of fixed points or of periodic points, then it is often called the *stable manifold*.

8.78 Definition. Let $\omega(x)$ be the forward limit of a point x . We say that $\omega(x)$ attracts y if $\omega(y) \subset \omega(x)$. The basin of attraction of $\omega(x)$ is the set of points which are attracted by $\omega(x)$. We say that $\omega(x)$ is an attractor if it attracts a substantial number of points, more precisely if it attracts a set of points of positive Lebesgue measure.

We then say that $\omega(x)$ is a chaotic set if the orbit of x is a chaotic orbit and $x \in \omega(x)$. Finally $\omega(x)$ is a chaotic attractor if $\omega(x)$ is both a chaotic set and an attractor.

8.79 Example. The map

$$f(x) = \frac{4}{\pi} \arctan x$$

has three fixed points $-1, 0$ and 1 . From Figure 8.26 we see that $-1, 1$ are sinks, while 0 is a source. The attractors are therefore $\{-1\}, \{1\}$, and their basins of attraction are respectively the negative half-line and the positive half-line.

8.80 Example. Consider the map in polar coordinates

$$f(r, \theta) := (r^2, \theta - \sin \theta) \quad r \geq 0, 0 \leq \theta < 2\pi.$$

It has three fixed points, $(0, 0)$, $(1, 0)$ and $(1, \pi)$. The origin and infinity are two attractors: iterations move every interval point of the unit disk to the origin and every extremal point to infinity. On the circle the dynamics moves all points but $(1, \pi)$, that is a source, to $(1, 0)$, see Figure 8.26.

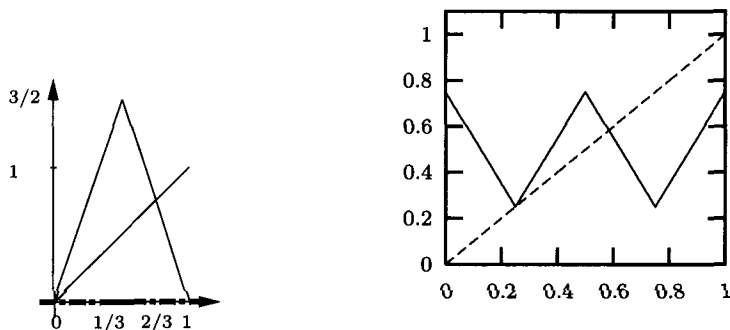


Figure 8.27. (a) The triangular map with slope 3. (b) The map W .

8.81 Example (The triangular map with slope 3). The following example shows that the basin of attraction of an attractor can be quite a complex set. Let us consider the triangular map $f: \mathbb{R} \rightarrow \mathbb{R}$ in Figure 8.27 (a). Clearly, the orbits with initial conditions either in $]-\infty, 0[$ and $]1, \infty[$ converge to $-\infty$. Also orbits with initial conditions in $]1/3, 2/3[$ converge to $-\infty$ since the first iteration maps this interval in $]1, \infty[$. It should then be clear that the basin of attraction of $-\infty$ under the map f is the complement in \mathbb{R} of the *Cantor middle-third set* defined below.

8.82 Example. Let us consider the map $f(r, \theta) := (r^{1/2}, 2\theta)$. The origin is a source and $\omega(0) = 0$. Taking into account that $(1, 2\theta)$ describes Bernoulli's shift on S^1 , we see that the unit circle is the forward limit set of an orbit with Lyapunov exponent $\log 2$ for almost all initial points in the circle, consequently the unit circle is a chaotic set. Finally all points but the origin are attracted to the unit circle; we therefore conclude that the unit circle is a *chaotic attractor*.

8.83 Example (The W map). Let us consider the piecewise linear map $f: [0, 1] \rightarrow [0, 1]$ in Figure 8.27 (b). Restricting our attention to the interval $[1/4, 3/4]$, the map acts as the triangular map with slope 1, while the points in $[0, 1] \setminus [1/4, 3/4]$ are mapped by the first iteration in $[1/4, 3/4]$. We therefore conclude that $[1/4, 3/4]$ is a chaotic attractor.

8.84 Example (The baker's map). Let us consider the area-preserving map given by

$$x_{n+1} = 2x_n \pmod{1}, \quad y_{n+1} = \begin{cases} 1/2 y_n & \text{if } 0 \leq x_n < 1/2, \\ 1/2 + 1/2 y_n & \text{if } 1/2 \leq x_n \leq 1, \end{cases}$$

and illustrated in Figure 8.28. Since the first component is Bernoulli's shift, we easily conclude that the entire square is a chaotic attractor.

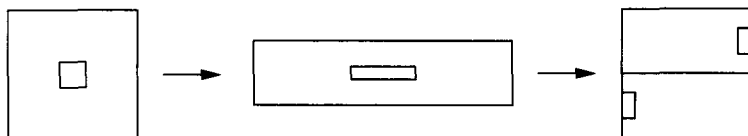


Figure 8.28. The baker's map.

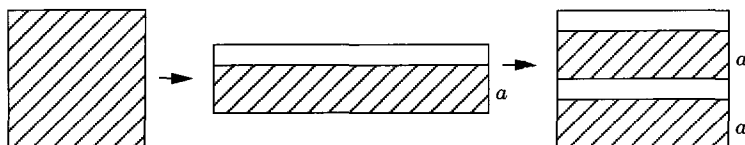


Figure 8.29. The dissipative baker's map.

8.85 Example (The baker's dissipative process). Consider now the process, similar to the baker's process,

$$x_{n+1} = 2x_n \pmod{1}, \quad y_{n+1} = \begin{cases} ay_n & \text{if } 0 \leq x_n < 1/2, \\ 1/2 + ay_n & \text{if } 1/2 \leq x_n \leq 1, \end{cases}$$

where $a < 1/2$, see Figure 8.29. This process is dissipative, that is, it does not preserve the area. The baker's dissipative process has still a *chaotic attractor* which is made now by a huge set of horizontal lines. As we shall see below, compare Example 8.99, this *strange attractor* is a *fractal*, in the sense that its "dimension" is strictly between 1 and 2.

8.86 Example. A process that is very similar to the baker's dissipative process is the one associated to the map

$$f(x, y) := \begin{cases} \left(\frac{1}{3}x, 2y\right) & \text{if } 0 < y \leq 1/2, \\ \left(\frac{1}{3}x + \frac{2}{3}, 2y - 1\right), & \text{if } 1/2 < y \leq 1. \end{cases} \quad (8.43)$$

The first iteration maps the square into the first and last third of the square, as shown in Figure 8.30. The figure shows also the second iteration. Clearly the attractor of the full square is again a chaotic attractor which is a *strange attractor*, being a set of lines which has a "noninteger dimension."

8.2.5 Cantor sets and other self-similar sets

a. Measure and dimension

8.87 Box counting dimension. A simple way to compute a "dimension" of a bounded subset of \mathbb{R}^2 is the following. Assume that A is contained in a rectangle R . Then divide each side of R in 2^k pieces, thus

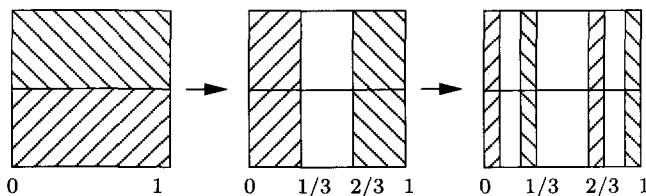


Figure 8.30. The first two iterations of the process in Example 8.86.

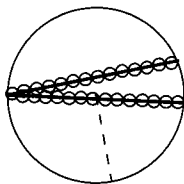


Figure 8.31.

dividing R into 4^k rectangles. Let $N(k)$ be the number of rectangles which touch A . If A is a line, then for k large $N(k)$ is of order 2^k , while if A has an interior part, then $N(k)$ is of order 4^k , thus it is reasonable to set the *box counting dimension* of A at scale 2^{-k} as

$$\frac{\log_2 N(k)}{k} = \frac{\log N(k)}{k \log 2}$$

and the *box counting dimension* as

$$\dim_B(A) := \lim_{k \rightarrow \infty} \frac{\log N(k)}{k \log 2} \quad (8.44)$$

provided the limit exists.

8.88 Hausdorff measure and Hausdorff dimension. Usually, dimension is associated to changes in measure under dilations: under a dilation in \mathbb{R}^n of factor λ , points do not change, segment and curve lengths are multiplied by λ , square and surface area are multiplied by λ^2 , while volumes are multiplied by λ^3 . So another possible definition of dimension goes through the definition of a measure that scales suitably under dilations.

There are many different measures in \mathbb{R}^n , $n \geq 1$, that are invariant under translations, yield the same measure for regular sets, and scale with a given power less than n under dilations. Among the many possibilities, the *Hausdorff measure*, introduced in 1918 by Felix Hausdorff (1869–1942), is sufficient and suited to our purposes.

Let us describe the Hausdorff measure \mathcal{H}^s in \mathbb{R}^n , $n \geq 1$. It is usual to define for any nonnegative real number $s \geq 0$,

$$\omega_s := \frac{\pi^{s/2}}{\frac{s}{2} \Gamma(\frac{s}{2})},$$

Γ being the gamma function, since, as one can show, for integral values of s , ω_s is the volume of the unit ball in \mathbb{R}^s .

8.89 Definition. Let $A \subset \mathbb{R}^n$. For any $\delta > 0$ we define

$$\mathcal{H}_\delta^s(A) := \inf \left\{ \sum_k \omega_s \left(\frac{\text{diam } C_k}{2} \right)^s \mid \bigcup_k C_k \supset A, \right. \\ \left. \{C_k\} \text{ are } n\text{-balls with } \text{diam } C_k < \delta, C_k \subset \mathbb{R}^n \right\}.$$

The spherical Hausdorff measure \mathcal{H}^s of A is then defined by

$$\mathcal{H}^s(A) := \lim_{\delta \rightarrow 0} \mathcal{H}_\delta^s(A), \quad \forall A \subset \mathbb{R}^n.$$

We notice that, \mathcal{H}^s is well defined, since \mathcal{H}_δ^s is increasing with δ ; moreover we notice that the naive definition

$$\widehat{\mathcal{H}}^s = \inf \left\{ \sum_k \omega_s \left(\frac{\text{diam } C_k}{2} \right)^s \mid A \subset \cup_k C_k \right\}$$

would not work, see Figure 8.31.

It is easily seen:

- (i) $\mathcal{H}^s(\lambda A) = \lambda^s \mathcal{H}^s(A)$, $\lambda > 0$, where $\lambda A = \{\lambda x \mid x \in A\}$,
- (ii) $\mathcal{H}^0(A) = \#A$ is the measure that counts the elements of A in \mathbb{R}^n
- (iii) if $s > n$, then $\mathcal{H}^s(A) = 0 \forall A \subset \mathbb{R}^n$,
- (iv) if $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is Lipschitz-continuous, then $\mathcal{H}^s(f(A)) \leq \text{Lip}(f)^s \mathcal{H}^s(A)$,
- (v) \mathcal{H}^s is invariant under translations and rotations.

Finally, we notice that

$$\mathcal{H}_\delta^r(A) \leq \left(\frac{\delta}{2} \right)^{r-s} \mathcal{H}_\delta^s(A), \quad \text{if } 0 \leq s < r,$$

that is, if $0 \leq s < r$, then

- (a) $\mathcal{H}^s(A) = +\infty$ if $\mathcal{H}^r(A) > 0$,
- (b) $\mathcal{H}^r(A) = 0$ if $\mathcal{H}^s(A) < \infty$.

In particular $\mathcal{H}^s(A)$ may be positive and finite only for one value of s .

8.90 Definition. The Hausdorff dimension of $A \subset \mathbb{R}^n$ is then defined as

$$\dim_{\mathcal{H}} A := \inf \left\{ s \geq 0 \mid \mathcal{H}^s(A) = 0 \right\}.$$

From the previous remarks, one easily infers

$$\dim_{\mathcal{H}} A = \sup \left\{ s \geq 0 \mid \mathcal{H}^s(A) = +\infty \right\}$$

and that $\dim_{\mathcal{H}} A = s$ if $0 < \mathcal{H}^s(A) < \infty$.

8.91 Definition. Sets in \mathbb{R}^n with nonintegral dimension are called fractals.

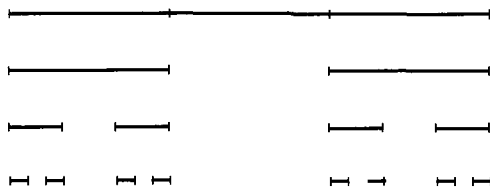


Figure 8.32. The sets E_0 , E_1 , E_2 and E_3 in the construction of the Cantor set $C_{1/3}$.

b. Cantor sets

Cantor sets are a family of subsets of \mathbb{R} among which the *Cantor middle-third set* is a prototype. They are obtained as follows. Choose $\delta \in]0, 1/2[$, ($\delta = 1/3$ for the Cantor middle-third), and set $E_0 = [0, 1]$. In the first step we define E_1 by removing from E_0 an open interval, centered at the middle point, of length $1 - 2\delta$. E_1 is then the union of two intervals of size δ . By induction, we define E_{k+1} by removing from each of the intervals of E_k a centered interval of length $\delta^k(1 - 2\delta)$. This way we get a decreasing sequence of sets $\{E_k\}$, E_k being the union of 2^k intervals of size δ^k . The *Cantor set associated to δ* is defined as

$$C_\delta := \bigcap_{k=0}^{\infty} E_k. \quad (8.45)$$

It is not difficult to show the following.

8.92 Proposition. *The Cantor middle-third set $C_{1/3}$, corresponding to $\delta = 1/3$, consists of all numbers in $[0, 1]$ that have a ternary expansion involving only the digits 0 and 2.*

Another way to look at Cantor sets is useful. Consider the two maps, actually two *contractive similitudes*, $S_1, S_2 : [0, 1] \rightarrow [0, 1]$ given by

$$S_1(x) := \delta x, \quad S_2(x) := \delta x + 1 - \delta,$$

and for any set $A \subset [0, 1]$, set $S(A) := S_1(A) \cup S_2(A)$.

Then observe that the sets $\{E_k\}$ in (8.45) are actually produced by the following dynamical system acting on sets,

$$\begin{cases} E_0 := [0, 1], \\ E_{k+1} := S(E_k), \end{cases}$$

i.e.,

$$E_k := \underbrace{S \circ S \circ \cdots \circ S}_k([0, 1]) =: S^k([0, 1])$$

from which, taking also into account that $E_{k+1} \subset E_k \forall k$, we infer

$$\begin{aligned}
C_\delta &= \bigcap_{k=1}^{\infty} E_k = \bigcap_{k=0}^{\infty} S(E_k) = (\text{since } E_{k+1} \subset E_k) \\
&= S\left(\bigcap_{k=0}^{\infty} E_k\right) = S(C_\delta).
\end{aligned} \tag{8.46}$$

8.93 Definition. A system $S := (S_1, S_2, \dots, S_N)$ of N contraction maps on \mathbb{R}^n is called an iterated function system, an IFS for short. A set $C \subset \mathbb{R}^n$ such that

$$C = \bigcup_{i=1}^N S_i(C), \quad \text{and} \quad S_i(C) \cap S_j(C) = \emptyset \quad \forall i \neq j$$

is called a self-similar set.

8.94 Remark. The terminology becomes clearer when (S_1, S_2, \dots, S_N) are *contractive similitudes*. Assuming for instance $N = 2$, and that S_1 and S_2 contract by a factor δ , conditions

$$C = S_1(C) \cup S_2(C), \quad S_1(C) \cap S_2(C) = \emptyset \tag{8.47}$$

say that C is the union of two pieces which are each a scaled down copy of C itself, that is C is *self-similar at the scale δ* . Moreover from (8.47) we also have

$$\begin{aligned}
C &= (S_1 \circ S_1(C)) \cup (S_1 \circ S_2(C)) \cup (S_2 \circ S_1(C)) \cup (S_2 \circ S_2(C)), \\
S_i \circ S_j(C) \cap S_h \circ S_k(C) &= \emptyset \quad \forall (i, j) \neq (h, k),
\end{aligned}$$

that is, C is also the union of four pieces that are scaled down versions of C of factor δ^2 . Proceeding by induction, for any k , C is also the union of 2^k pieces each of which is a scaled down copy of C by a factor δ^k . This is self-similarity.

8.95 . A more explicit description of the Cantor set C_δ is the following one. If the *base points* of the Cantor set are defined by induction by

$$b_{0,1} = 0, \quad b_{k+1,j} = \begin{cases} \delta b_{k,j} & \text{if } j = 1, \dots, 2^k, \\ 1 - \delta + \delta b_{k,j} & \text{if } j = 2^k + 1, \dots, 2^{k+1} \end{cases}$$

then

$$I_{k-1,j} := b_{k-1,j} + \delta^{k-1}(\delta, 1 - \delta), \quad j = 1, \dots, 2^{k-1},$$

are the intervals to be deleted from E_{k-1} to get E_k at the k -step, and

$$J_{k,j} := b_{k,j} + \delta^k[0, 1], \quad j = 1, \dots, 2^k,$$

are the intervals whose union is E_k . We have set $\alpha[a, b] := [\alpha a, \alpha b]$. Consequently

$$C_\delta = \bigcap_{k=0}^{\infty} \left(\bigcup_{j=1}^{2^k} J_{j,k} \right).$$



Figure 8.33. Helge von Koch (1870–1924) and Wacław Sierpinski (1882–1969).

This shows again that C is *self-similar*, and more precisely that

$$b_{k,j} + \delta^k C_\delta = C_\delta \cap J_{k,j}. \quad (8.48)$$

In fact,

$$x \in J_{h,j} \quad \text{if and only if} \quad \delta^k x \in J_{h+k,j}, \quad \forall h, k, j,$$

hence

$$\delta^k E_h = E_{h+k} \cap [0, \delta^k], \quad \text{i.e.,} \quad b_{k,j} + \delta^k E_h = E_{h+k} \cap J_{k,j},$$

for all $h, k \geq 0$, $j = 1, \dots, 2^k$; and (8.48) follows by taking into account the intersection on h .

c. Iterated function systems

Self-similarity is even more evident in the two-dimensional Cantor set, known also as *Sierpinski's square*, obtained by dividing the unit square in 3^n squares and removing the central square from each of the squares left at the n -th, see Figure 8.34, or as the so-called *Sierpinski's carpet*, obtained by removing the central cross (see, for example, Figure 8.37), or in the *Sierpinski's gasket* (see Figure 8.35).

Another “1-dimensional” example is the *von Koch curve*; compare Figure 6.23. The curve of von Koch, contrary to regular curves and similarly to the examples above, has the property that any enlargement or blow-up

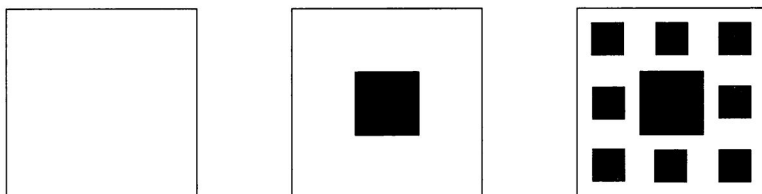


Figure 8.34. The first steps in the construction of Sierpinski's square.

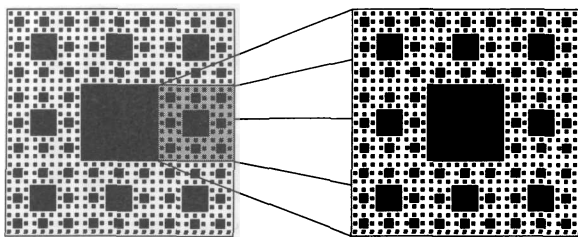


Figure 8.35. Another step in the construction of Sierpinski's square.

of it does not simplify, instead it leaves the complex structure unchanged. This may be presented as an informal definition of self-similarity.

All these examples are defined following the same pattern. One starts with an IFS system of contractions (S_1, \dots, S_N) on \mathbb{R}^n , and, as proved by Felix Hausdorff (1869–1942), one can show (but we shall not do it here) that there is a unique nonvoid, bounded and closed set C such that

$$C = \cup_{i=1}^N S_i(C).$$

We refer to it as to *the invariant set* of the IFS (S_1, S_2, \dots, S_N) . C is in fact a fixed point for the map $A \rightarrow \cup_{i=1}^N S_i(A)$ on the so-called *Hausdorff space*. Moreover, introducing a suitable notion of “distance between sets,” C can be found as the *limit* of the sequence of sets $\{F_k\}$ defined by

$$\begin{cases} F_0 \text{ an arbitrary closed bounded and nonvoid set,} \\ F_{k+1} := \cup_{i=1}^N S_i(F_k). \end{cases}$$

The reader will recognize the iterative procedure as the same defining Cantor sets $C_\sigma \subset \mathbb{R}$.

In general the invariant set is not self-similar, but it is under suitable sufficient conditions.

d. Dimension of the invariant set

We restrict ourselves to an IFS S_1, S_2, \dots, S_N of *contractive similitudes*

$$|S_i(x) - S_i(y)| = L_i |x - y| \quad \forall x, y \in \mathbb{R}^n, i = 1, \dots, N.$$

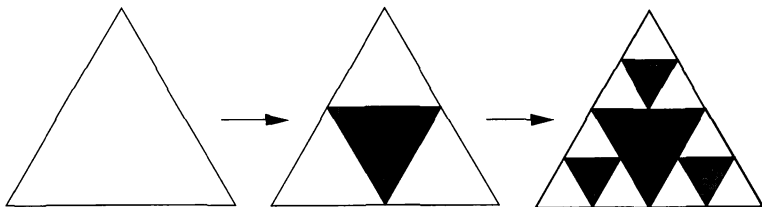


Figure 8.36. The first steps in the construction of Sierpinski's gasket.

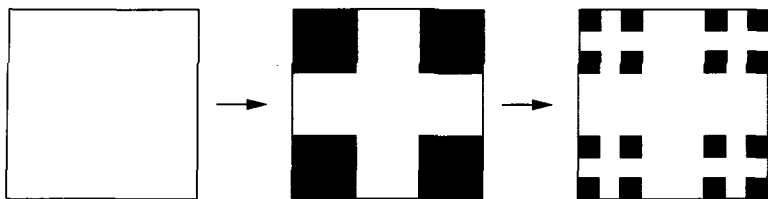


Figure 8.37. The first steps in the construction of Sierpinski's carpet.

We define the *geometric dimension* of that IFS as the unique nonnegative real number D such that

$$\sum_{i=1}^N L_i^D = 1. \quad (8.49)$$

In general the geometric dimension has no relation with the more geometric definitions of dimension. However, we have the following.

8.96 Proposition. *If the invariant set C of an IFS of contractive similitudes is self-similar and for some s we have $0 < \mathcal{H}^s(C) < +\infty$, then s is the geometric dimension of the IFS.*

Proof. In fact,

$$\mathcal{H}^s(C) = \sum_{i=1}^N \mathcal{H}^s(S_i(C)) = \mathcal{H}^s(C) \sum_{i=1}^N L_i^s,$$

i.e., $\sum_{i=1}^N L_i^s = 1$, since $0 < \mathcal{H}^s(C) < \infty$. □

Of course, going in the opposite direction is more useful. We state without proof the following theorem which gives a full description of some IFSs of contractive similitudes.

8.97 Definition. *One says that an IFS of contractive similitudes satisfies the open set condition if there exists an open set $\Omega \subset \mathbb{R}^n$ such that*



Figure 8.38. Three iterates going to von Koch's curve starting with the segment of vertices $(0, 0)$ and $(1, 0)$.

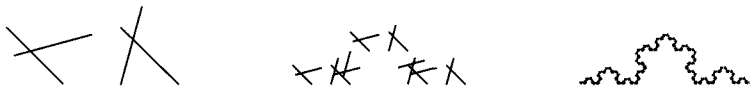


Figure 8.39. The second, third, and fifth iterate going to von Koch's curve starting with the segment of vertices $(0, 1)$ and $(1, 0)$.

$$\begin{cases} \bigcup_i S_i(\Omega) \subset \Omega, \\ S_i(\Omega) \cap S_j(\Omega) = \emptyset \quad \forall i \neq j. \end{cases}$$

8.98 Theorem. Let (S_1, S_2, \dots, S_N) be an IFS of contractive similitudes which contract respectively of L_1, \dots, L_N , let C be the invariant set of the IFS, and let d be the geometrical dimension of the IFS, $\sum_{i=1}^N L_i^d = 1$.

If the IFS satisfies the open set condition, then

- (i) $0 < \mathcal{H}^d(C) < +\infty$, hence $\dim_{\mathcal{H}}(C) = d$,
- (ii) $\mathcal{H}^d(S_i(C) \cap S_j(C)) = 0 \quad \forall i \neq j$.

Moreover, each piece of C has the same dimension, and d is also the box-counting dimension of C .¹⁴

Theorem 8.98 yields a way to conclude that the invariant set of the IFS of contractive similitudes that satisfy the open set condition is essentially self-similar since C is a union of N scaled down copies of C itself that can overlap but in a nonessential way, as the intersections have zero measure. As a by-product we have a formula, the geometric dimension, to compute effectively the dimension of C .

8.99 Example (Cantor set in \mathbb{R}). As we have seen, this is the invariant set of the two contractive similitudes of \mathbb{R} ,

$$S_1(x) = \delta x, \quad S_2(x) = \delta x + 1 - \delta.$$

(S_1, S_2) satisfies the open set condition, Ω being the open interval $]0, 1[$. Therefore the Cantor set is self-similar and by Theorem 8.98 has dimension d given by

$$2\delta^d = 1, \quad \text{i.e.,} \quad d = d(\delta) = \frac{\log 2}{\log(1/\delta)}.$$

Notice that $0 < d(\delta) < 1$ being that $0 < \delta < 1/2$. See Figure 8.32 for the first iterations starting from the segment $[0, 1]$.

¹⁴ See, e.g., J. E. Hutchinson, *Fractals and self-similarity*, Indiana Univ. Math. J. 30, 713–747 (1981).

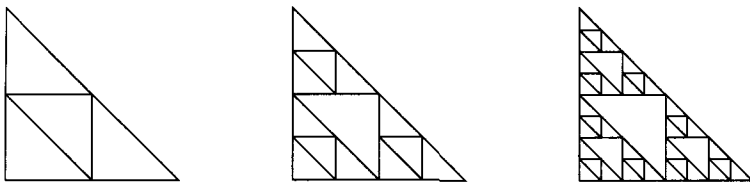


Figure 8.40. The first three iterates going to Sierpinski's gasket starting from the triangle of vertices $(0, 0)$, $(1, 0)$ and $(0, 1)$.

8.100 Example (Sierpinski's gasket). This is the invariant set for the IFS of contractive similitudes of \mathbb{R}^2 defined by

$$\begin{aligned} S_1 \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} x/2 \\ y/2 \end{pmatrix} + \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}, \\ S_2 \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} x/2 \\ y/2 \end{pmatrix} + \begin{pmatrix} 0 \\ 1/2 \end{pmatrix}, \\ S_3 \begin{pmatrix} x \\ y \end{pmatrix} &= \begin{pmatrix} x/2 \\ y/2 \end{pmatrix} + \begin{pmatrix} 1/2 \\ 0 \end{pmatrix}. \end{aligned}$$

(S_1, S_2, S_3) satisfies the open set condition, Ω being the open triangle of vertices $(0, 0)$, $(0, 1)$ and $(1, 0)$. By Theorem 8.98 Sierpinski's gasket is essentially self-similar and has a nonintegral dimension d given by

$$3\left(\frac{1}{2}\right)^d = 1, \quad \text{i.e.,} \quad d = \frac{\log 3}{\log 2} > 1.$$

See Figure 8.40 for the first iterations starting from the triangle E_0 of vertices $(0, 0)$, $(0, 1)$ and $(1, 0)$.

8.101 Example (Sierpinski's square). This is the invariant set for the IFS of eight contractive similitudes of \mathbb{R}^2 defined by

$$S_i \begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{3} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a_i \\ b_i \end{pmatrix}$$

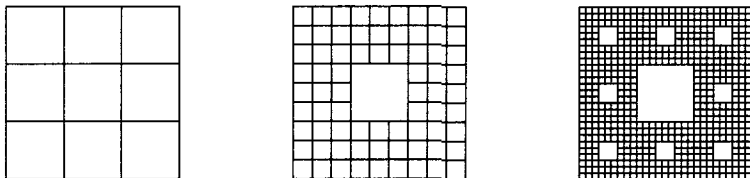


Figure 8.41. The first three iterates going to Sierpinski's square starting from the square of vertices $(0, 0)$, $(1, 0)$, $(1, 1)$ and $(0, 1)$.

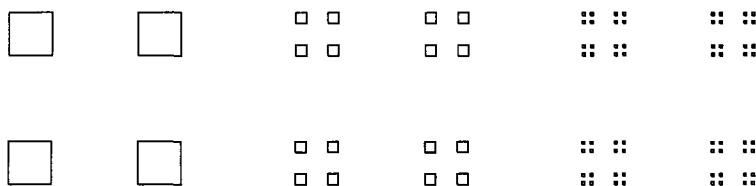


Figure 8.42. The first three iterates going to a Sierpinski's carpet that contracts to $1/4$ starting from the square of vertices $(0,0)$, $(1,0)$, $(1,1)$ and $(0,1)$.

where (a_i, b_i) is one of $(0,0)$, $(1/3,0)$, $(2/3,0)$, $(0,2/3)$, $(1/3,2/3)$, $(2/3,2/3)$, $(0,1/3)$, $(2/3,1/3)$. (S_1, S_2, \dots, S_8) satisfies the open set condition, Ω being the open square of vertices $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$. By Theorem 8.98 Sierpinski's square is essentially self-similar and has a nonintegral dimension d given by

$$8\left(\frac{1}{3}\right)^d = 1, \quad \text{i.e.,} \quad d = \frac{\log 8}{\log 3}.$$

Notice that $1 < d < 2$. See Figure 8.41 for the first iterations starting from the rectangle E_0 of vertices $(0,0)$, $(0,1)$, $(1,1)$ and $(1,0)$.

8.102 Example (Sierpinski's carpet). This is the invariant set for the IFS of four contractive similitudes of \mathbb{R}^2 defined by

$$S_i \begin{pmatrix} x \\ y \end{pmatrix} = q \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a_i \\ b_i \end{pmatrix}$$

where $q < 1/2$ and (a_i, b_i) is one of $(0,0)$, $(1-q,0)$, $(0,1-q)$ and $(1-q,1-q)$. (S_1, S_2, \dots, S_4) satisfies the open set condition, Ω being the open square of vertices $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$. By Theorem 8.98 Sierpinski's carpet is essentially self-similar and has dimension d given by

$$4q^d = 1, \quad \text{i.e.,} \quad d = \frac{\log(1/4)}{\log q}.$$

Notice that for $q = 1/4$, we get a set of dimension 1. See Figure 8.42 for the first iterations starting from the rectangle E_0 of vertices $(0,0)$, $(0,1)$, $(1,1)$ and $(1,0)$.

8.103 Example (Snowflake). This is the invariant set for the IFS of nine contractive similitudes of \mathbb{R}^2 defined by

$$S_1 \begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{2} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 1/4 \\ 1/4 \end{pmatrix},$$

and for $i = 2, \dots, 9$,

$$S_i \begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{8} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a_i \\ b_i \end{pmatrix}$$

where (a_i, b_i) is one of $(0,0)$, $(0,7/8)$, $(7/8,0)$, $(7/8,7/8)$, $(1/8,1/8)$, $(1/8,3/4)$, $(3/4,1/8)$, $(3/4,3/4)$. (S_1, S_2, \dots, S_9) satisfies the open set condition, Ω being the open square of

vertices $(0, 0)$, $(0, 1)$, $(1, 0)$ and $(1, 1)$. By Theorem 8.98 the snowflake is essentially self-similar and has dimension d given by

$$1 \cdot \left(\frac{1}{2}\right)^d + 8\left(\frac{1}{8}\right)^d = 1, \quad \text{i.e.,} \quad d = 1.25996 \dots$$

See Figure 8.43 for the first iterations starting from the rectangle E_0 of vertices $(0, 0)$, $(0, 1)$, $(1, 1)$ and $(1, 0)$.

8.104 Example (von Koch's curve). This is the invariant set for the IFS of contractive similitudes of \mathbb{R}^2 defined by

$$\begin{aligned} S_1 \begin{pmatrix} x \\ y \end{pmatrix} &= \frac{1}{3} \begin{pmatrix} x \\ y \end{pmatrix}, \\ S_2 \begin{pmatrix} x \\ y \end{pmatrix} &= \frac{1}{3} R\left(\frac{\pi}{3}\right) \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 0 \\ 1/3 \end{pmatrix}, \\ S_3 \begin{pmatrix} x \\ y \end{pmatrix} &= \frac{1}{3} R\left(-\frac{\pi}{3}\right) \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 1/2 \\ \sqrt{3}/2 \end{pmatrix}, \\ S_4 \begin{pmatrix} x \\ y \end{pmatrix} &= \frac{1}{3} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} 2/3 \\ 0 \end{pmatrix}, \end{aligned}$$

where $R(\theta)$ denotes the rotation matrix by an angle θ measured counterclockwise

$$R(\theta) = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}.$$

(S_1, S_2, S_3, S_4) satisfies the open set condition, Ω being the open triangle of vertices $(0, 0)$, $(0, 1)$ and $(1/2, \sqrt{3}/2)$. By Theorem 8.98 the von Koch curve is essentially self-similar and has a nonintegral dimension d given by

$$4\left(\frac{1}{3}\right)^d = 1, \quad \text{i.e.,} \quad d = \frac{\log 4}{\log 3} > 1.$$

See Figure 8.38 for the first iterations starting from the segment $[0, 1]$ on the real axis.

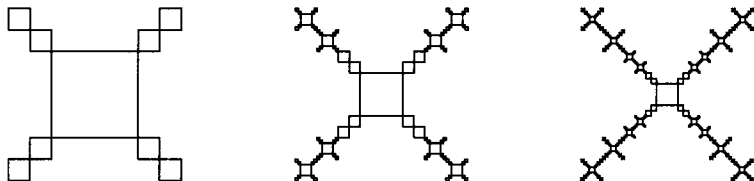


Figure 8.43. The first three iterates going to snowflake starting from the square of vertices $(0, 0)$, $(1, 0)$, $(1, 1)$ and $(0, 1)$.

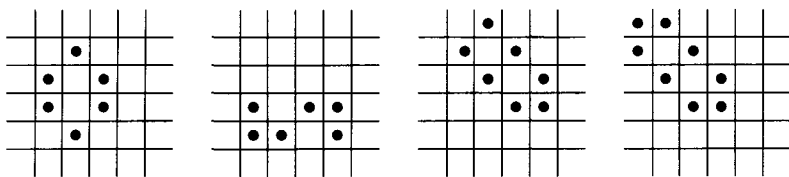


Figure 8.44. Some stable configurations: beehive, snake, long boat, longship.

8.3 Two-Dimensional Dynamical Systems

Of course, we have no chance here to discuss multidimensional discrete processes. We confine ourselves to commenting two quite popular processes: the *game of life* and the *dynamics of complex maps*

$$P_c(z) := z^2 + c. \quad (8.50)$$

8.3.1 Game of life

The game of life is a well-known dynamical system that was conceived in the 1960s by John Conway in Cambridge and has attracted many people, especially biologists. A comprehensive presentation of it can be found in E. R. Berlekamp, J. H. Conway, and R. K. Guy, *Winning Ways*, New York, 1982. Therefore, here we confine ourselves to saying a few words about it.

Imagine that the plane is decomposed into square cells, and that each of these cells can be left vacant or can be filled with a black disc. A *state* of our system is a distribution of black discs into a finite number of cells. The denumerable family of all states is denoted by X . The transition law from x to $T(x)$ is defined by applying successively the following three rules to x :

- (i) *two or three neighbors keep you alive*. A cell that is occupied in the state x will be occupied in $T(x)$ if and only if it has two or three neighbors that are occupied in the state x .

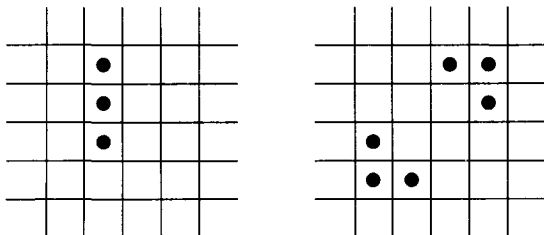


Figure 8.45. Two 2-periodic states: blinker and beacon.

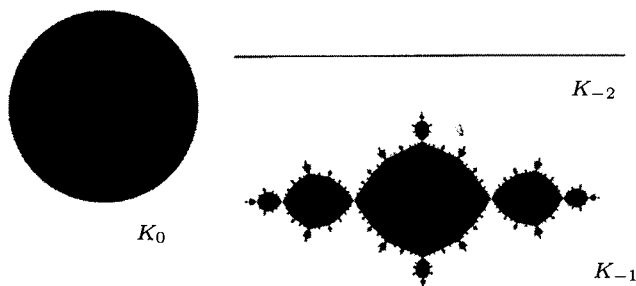


Figure 8.46. The Julia sets of K_0 , K_{-1} , K_{-2} .

- (ii) *three neighbors create life.* A cell that is vacant in the state x will be occupied in the state $T(x)$ if and only if it has precisely three neighbors that are occupied in x .
- (iii) *you die if you are alone or in the crowd.* If neither (i) nor (ii) apply to a given cell, this cell will be vacant in the state $T(x)$.

Figure 8.44 shows some of the fixed points of T , while Figure 8.45 shows two 2-periodic states.

But the game of life has a very rich dynamics. For instance, one can show:

- (i) the game of life can simulate any computer,
- (ii) there exists at least one garden of Eden, i.e., a configuration that has no predecessor.

8.3.2 Fractal boundaries

Let us discuss now very briefly some of the dynamics of the maps (8.50).

As we saw, when restricted to the real case, they are conjugate to the logistic maps. The complexity of the maps in (8.50) in \mathbb{C} therefore may better motivate the complexity of the logistic map.

Let us begin with the case $c = 0$. The map has a sink at $z = 0$ with basin of attraction the unit disc $\{z \mid |z| < 1\}$. Points of $S^1 := \{z \mid |z| = 1\}$ are mapped into S^1 with double argument, while the orbit of any exterior point z , $|z| > 1$, diverges to infinity.

Quite more complicated is the case $c \neq 0$. The study of the iterates of complex maps begins with the works of Pierre Fatou (1878–1929) and Gaston Julia (1893–1978). With reference to the maps $P_c(z) = z^2 + c$, a natural question to ask is: which points in \mathbb{C} have unbounded orbits?

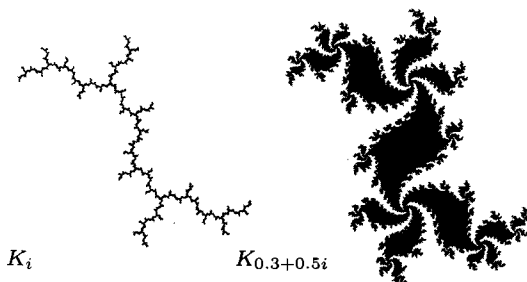


Figure 8.47. The Julia sets of K_i and $K_{0.3+0.5i}$.

a. Julia sets

8.105 Definition. We denote by J_c the set of points in \mathbb{C} which have bounded orbits

$$K_c := \{z \mid P_c^n(z) \not\rightarrow \infty\}.$$

The boundary¹⁵ P_c of K_c is called a Julia set.

8.106 ¶. As we have seen $K_0 = \{z \mid |z| \leq 1\}$. Show then that $K_{-2} = [-2, 2]$.

The Julia sets corresponding to $c = 0, -2$ are the only ones that are geometrically simple; for all other values of c the corresponding Julia sets are fractal.

The following theorems, that we state without proof, are due to Fatou and Julia. The first shows that the dynamics is fairly controlled by critical points.

8.107 Theorem (Fatou). Every attracting cycle of a polynomial map P attracts at least one critical point.

For instance, a quadratic polynomial has infinitely many periodic circles, however at most one may be attractive, as there can only be one attractive critical point. The map $P_{-i} := z^2 - i$ has a repelling period 2 point, as $P_{-i}^2(i) = -1 - i$, consequently it has no attractor.

8.108 Theorem (Julia). Let $K_c := \{z \mid P_c^k(z) \not\rightarrow \infty\}$. Then

- (i) K_c is connected¹⁶ if and only if the origin belongs to K_c ,
- (ii) K_c is a Cantor type set if the origin does not belong to K_c .

¹⁵ A point z is in the boundary of K_c if in every ball centered at z there are both points of K_c and of $\mathbb{C} \setminus K_c$.

¹⁶ We recall that a set in \mathbb{C} is connected if any two of its points can be joined by a continuous curve that is in the set.

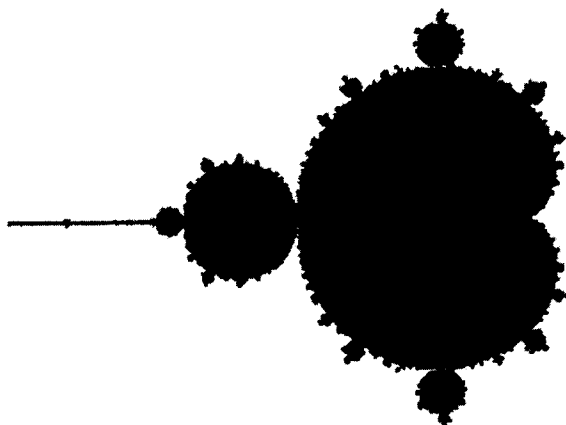


Figure 8.48. (An approximation of) Mandelbrot's set.

b. Mandelbrot set

In consideration of the relevance of the orbit of the origin,

8.109 Definition. We define the Mandelbrot set as

$$M := \left\{ c \mid 0 \text{ is not in the basin of attraction of } P_c \right\}.$$

8.110 ¶. Show that $0, -i \in M$, while $-1 \notin M$.

It is worth noticing that Julia sets for P_c are never empty and that the Mandelbrot set is extremely complex, as it is connected and has a fractal boundary.

8.3.3 Fractals on the computer

Julia and Mandelbrot sets exert a tremendous esthetic fascination when represented on the screen of a computer, and, probably for this reason, they have become very popular.

8.111 The sets K_c . Given c , we can visualize (approximately) the orbits of each z as follows. We choose a grid of points in the plane and compute a fixed number \bar{k} (say 100, 300 or 1000) of iterates of each point of the grid. If the iterates $P_c^k(z)$, $k \leq \bar{k}$, remain bounded, $|P_c^k(z)| \leq |c| + 1$, we colour z in black; if the orbit becomes unbounded, $|P_c^k(z)| > |c| + 1$ for some k , we colour z in white. Notice, in fact, that if $|P_c^k(z)| > |c| + 1$ for some k , then $|P_c^k(z)| \rightarrow \infty$ as $k \rightarrow \infty$.

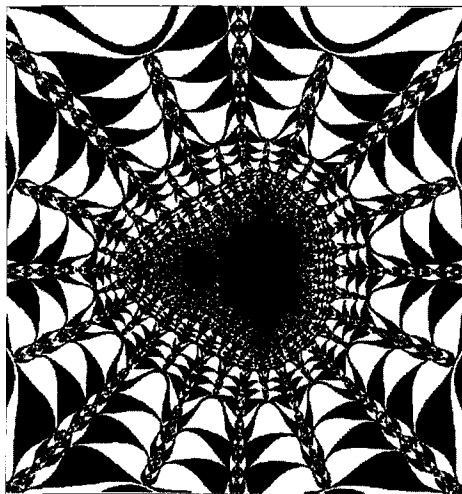


Figure 8.49. A picture of the Mandelbrot set by a computer program.

8.112 Mandelbrot set. In order to visualize the Mandelbrot set, we proceed similarly. But now the points in the grid refer to c and the initial value of the orbit is always zero.

8.113 Julia and Mandelbrot sets in colour. If $c \notin M$, then $P_c^n(0) \rightarrow \infty$ as $n \rightarrow \infty$. We colour the point c in the grid according to the number of iterates needed to leave a disk of prescribed radius R . We can proceed similarly for Julia sets.

8.4 Exercises

8.114 ¶. Show that

$$\sqrt{6 + \sqrt{6 + \sqrt{6 + \cdots}}} = 3, \quad \sqrt{1 + \sqrt{1 + \sqrt{1 + \cdots}}} = 1 + \frac{1}{1 + \frac{1}{1 + \frac{1}{1 + \cdots}}}.$$

8.115 ¶. Discuss the recurrence

$$\begin{cases} x_0 = \alpha > 0, \\ x_{n+1} = \frac{1}{x_n^2} \end{cases} \quad n \geq 0.$$

8.116 ¶ Campanato's lemma. Let $\phi :]0, 1] \rightarrow \mathbb{R}$ be an increasing function such that

$$\phi(r) \leq A\left(\frac{r}{R}\right)^\alpha \phi(R) + BR^\beta \quad \text{for all } 0 < r, R \leq 1,$$

where A, B, α, β are nonnegative constants and $0 < \alpha < \beta$. Show that

$$\phi(r) \leq Cr^\beta,$$

C being a constant depending only on $\alpha, \beta, \phi(R)$.

8.117 ¶ A useful lemma. Let $\phi:]0, 1] \rightarrow \mathbb{R}$ be an increasing function such that

$$\phi(r) \leq A(R-r)^{-\alpha} + B + \theta \phi(R) \quad \text{for all } 0 < r, R \leq 1,$$

where A, B, α, θ are nonnegative constants, $0 < \alpha$ and $0 \leq \theta < 1$. Show that

$$\phi(r) \leq C\left(A(R-r)^{-\alpha} + B\right),$$

C being a constant depending only on α, θ . [Hint: Apply the assumption to $r = r_n$, $R = r_{n+1}$, $\{r_n\}$ being a suitable increasing sequence that converges geometrically to R .]

8.118 ¶. Many identities involving Fibonacci numbers $\{f_n\}$ are known. The following exercises list some of them. Show the following.

- (i) $\sum_{j=1}^n f_j = f_{n+2} - 1$.
- (ii) $\sum_{j=1}^n f_j^2 = f_n f_{n+1}$.
- (iii) CASSINI IDENTITY. $f_{n-1} f_{n+1} - f_n^2 = (-1)^n$.
- (iv) $\sum_{j=0}^n \binom{n-j}{j} = f_n$.
- (v) CESÀRO. $\sum_{j=0}^n \binom{n}{j} f_j = f_{2n}$.
- (vi) LUCAS. g.c.d. $(f_p, f_q) = f_{\text{g.c.d.}(p, q)}$.
- (vii) For all $n \geq 1$, the numbers $f_n^2 + f_{n+1}^2$ and $f_{n+1}^2 - f_{n-1}^2$ are Fibonacci numbers.
- (viii) $\sum_{j=1}^{2n-1} f_j f_{j+1} = f_{2n}^2$.
- (ix) $f_{n+1}/f_n \rightarrow \tau := (1 + \sqrt{5})/2$.
- (x) $\tau^n = \tau f_n + f_{n-1}$ where $\tau = (1 + \sqrt{5})/2$.
- (xi) $\sum_{j=1}^{\infty} \frac{1}{f_j f_{j+1}} = 1$.
- (xii) $\sum_{j=1}^{\infty} (-1)^{j-1} \frac{1}{f_j f_{j+1}} = \tau^{-2}$, $\tau := (1 + \sqrt{5})/2$.
- (xiii) $\sum_{j=1}^{\infty} \frac{1}{f_j} = 4 - \tau$, $\tau := (1 + \sqrt{5})/2$.

8.119 ¶. Let $\{x_n\}$ be the Heaviside sequence, $x_n = 1 \forall n$. Show that $\mathcal{Z}\{x\}(z) = z/(z-1)$.

8.120 ¶. Let $\{x_n\}$ be the linear increasing sequence, $x_n = an \forall n$. Show that $\mathcal{Z}\{x\}(z) = az/(z-1)^2$.

8.121 ¶. Suppose that the \mathcal{Z} -transform of a sequence $a = \{a_n\}$ is a rational function near infinity, $\mathcal{Z}\{a\}(z) = A(z)/B(z)$, $A(z), B(z)$ being polynomials. Find the sequence $a = \{a_n\}$ in terms of A and B . [Hint: Use the Hermite decomposition formula.]

8.122 ¶. Let $\{x_n\}$ be the impulse sequence

$$x_n := (\underbrace{0, \dots, 0}_h, \underbrace{1, \dots, 1}_k, 1, 1, \dots).$$

Show that $\mathcal{Z}\{x\}(z) = \frac{1}{z^{h+k-1}} \frac{z^k - 1}{z - 1}$.

8.123 ¶. Let $\{x_n\}$ be a sequence and let $k \geq 1$. Define the sequence $m := \{m_n\}$ to be the sequence of traveling k -means by

$$\begin{aligned} m_0 &:= x_0, \\ m_1 &:= \frac{x_0 + x_1}{2}, \\ &\dots \\ m_{k-1} &:= \frac{x_0 + x_1 + \dots + x_{k-1}}{k} \end{aligned}$$

and for any $n \geq k$,

$$m_n := \frac{x_{n-k+1} + x_{n-k+2} + \dots + x_n}{k}.$$

Compute $\mathcal{Z}\{m\}(z)$.

8.124 ¶. Let $\{x_n\}$ be the orbit of a dynamical system governed by a second order difference equation. Show that the orbits are asymptotically stable if both the roots of the characteristic equation λ_1, λ_2 satisfy $|\lambda_1|, |\lambda_2| < 1$. Show that the system is stable if and only if either $|\lambda_1|, |\lambda_2| < 1$ or $|\lambda_1| = |\lambda_2| = 1$ with $\lambda_1 \neq \lambda_2$.

8.125 ¶¶. Let R be a rectangle of sides 1 and $h < 1$. From R we cut a square of side h and we are left with a rectangle of side h and $1 - h$. Then we reiterate the procedure. Under what conditions on h will the process never end?

8.126 ¶. Assuming that for $|z| < 1$,

$$\frac{e^z}{1-z} = a_0 + a_1 z + a_2 z^2 + \dots,$$

show that $a_n = \sum_{k=0}^n \frac{1}{k!}$. [Hint: Notice that $a_n - a_{n-1} = 1/n!$]

8.127 ¶. Assuming that a solution of

$$\begin{cases} (6x^2 - 5x + 1)y'' + 2(12x - 5)y' + 12y = 0, \\ y(0) = 1, \\ y'(0) = 0 \end{cases}$$

can be written as $\sum_{n=0}^{\infty} a_n x^n$, find the a_n . [Hint: Show that $a_{n+2} - 5a_{n+1} + 6a_n = 0 \forall n$.]

8.128 ¶. Check that

$$\frac{\log(1+z)}{1-z} = \sum_{n=0}^{\infty} a_n z^n \quad \text{with} \quad a_n = 1 + \sum_{k=1}^{n-1} \frac{(-1)^k}{n+1}.$$

8.129 ¶. Study fixed points of the maps

$$f(x) = \frac{1}{3}x^3 + \frac{2}{3}, \quad f(x) = x^4 - 3x^2 + 3x.$$

8.130 ¶. In dependence on the initial value x_0 , discuss the behavior at ∞ of

$$\begin{aligned} x_{k+1} &= \frac{x_k}{3} + \frac{4}{3}, & x_{k+1} &= 2x_k e^{-x_k/2}, \\ x_{k+1} &= x_k - x_k^2, & x_{k+1} &= \frac{1}{2 - x_k}, \\ x_{k+1} &= 1 - x_k + x_k^2, & x_{k+1} &= \frac{x_k}{1 + \sqrt{1 + x_k^2}}. \end{aligned}$$

8.131 ¶¶. Show that the fractional part of $(\frac{1+\sqrt{5}}{2})^n$ is not equidistributed, since

$$\frac{\#\{n \mid 1 \leq n \leq N, (\frac{1+\sqrt{5}}{2})^n \in [1/4, 3/4]\}}{N} \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

[Hint: Use that $u_n := ((1 + \sqrt{5})/2)^n + ((1 - \sqrt{5})/2)^n$ is an integer.]

A. Mathematicians and Other Scientists

Niels Henrik Abel (1802–1829)
Abu al Khwarizmi (790–850)
Archimedes of Syracuse (287BC–212BC)
Jean Argand (1768–1822)
Aristotle (384BC–322BC)
Antoine Arnauld (1612–1694)
Vladimir Arnold (1937–)
Eric Temple Bell (1883–1960)
George Berkeley (1685–1753)
Jacob Bernoulli (1654–1705)
Johann Bernoulli (1667–1748)
Felix Bernstein (1878–1956)
Wilhelm Bessel (1784–1846)
Enrico Betti (1823–1892)
Luigi Bianchi (1856–1928)
George Birkhoff (1884–1944)
Bernhard Bolzano (1781–1848)
Rafael Bombelli (1526–1573)
Emile Borel (1871–1956)
Satyendranath Bose (1894–1974)
Charles Brianchon (1783–1864)
L. E. Brouwer (1881–1966)
Cesare Burali-Forti (1861–1931)
Sergio Campanato (1930–)
Georg Cantor (1845–1918)
Girolamo Cardano (1501–1576)
Robert Carmichael (1879–1967)
Eugène Catalan (1814–1894)
Augustin-Louis Cauchy (1789–1857)
Ernesto Cesàro (1859–1906)
Paul Cohen (1934–)
Jean d'Alembert (1717–1783)
Abraham de Moivre (1667–1754)
Richard Dedekind (1831–1916)
René Descartes (1596–1650)

Ulisse Dini (1845–1918)
Diophantus of Alexandria (200–284)
Paul Dirac (1902–1984)
Lejeune Dirichlet (1805–1859)
Pierre Duhem (1861–1916)
Albert Einstein (1879–1955)
Gotthold Eisenstein (1823–1852)
Federigo Enriques (1871–1946)
Eratosthenes of Cyrene (276BC–197BC)
Paul Erdős (1913–1996)
Euclid of Alexandria (325BC–265BC)
Eudoxus of Cnidus (408BC–355BC)
Leonhard Euler (1707–1783)
Giulio Fagnano (1682–1766)
Pierre Fatou (1878–1929)
Mitchell Feigenbaum (1944–)
Pierre de Fermat (1601–1665)
Enrico Fermi (1901–1954)
Scipione del Ferro (1465–1526)
Karl Feuerbach (1800–1834)
Leonardo Pisano (1170–1250),
called Fibonacci
Joseph Fourier (1768–1830)
Abraham A. Fraenkel (1891–1965)
Gottlob Frege (1848–1925)
Galileo Galilei (1564–1642)
Evariste Galois (1811–1832)
Carl Friedrich Gauss (1777–1855)
Kurt Gödel (1906–1978)
James Gregory (1638–1675)
Jacques Hadamard (1865–1963)
William R. Hamilton (1805–1865)
G. H. Hardy (1877–1947)

- Felix Hausdorff (1869–1942)
 Oliver Heaviside (1850–1925)
 Charles Hermite (1822–1901)
 Heron of Alexandria (IAD)
 Thomas Herriot (1560–1621)
 David Hilbert (1862–1943)
 Christiaan Huygens (1629–1695)
 James Ivory (1765–1842)
 Camille Jordan (1838–1922)
 Gaston Julia (1893–1978)
 Helge von Koch (1870–1924)
 Andrey Kolmogorov (1903–1987)
 Leopold Kronecker (1823–1891)
 Martin Kutta (1867–1944)
 Joseph-Louis Lagrange (1736–1813)
 Pierre-Simon Laplace (1749–1827)
 Adrien-Marie Legendre (1752–1833)
 Gottfried von Leibniz (1646–1716)
 Carl von Lindemann (1852–1939)
 Hendrik Lorentz (1853–1928)
 Alfred Lotka (1880–1946)
 F. Edouard Lucas (1842–1891)
 Aleksandr Lyapunov (1857–1918)
 Colin MacLaurin (1698–1746)
 Francesco Maurolico (1494–1575)
 Pietro Mengoli (1626–1686)
 Frank Morley (1860–1937)
 Jurgen Moser (1928–1999)
 Sir Isaac Newton (1643–1727)
 Nicomachus of Gerasa (60AD–120)
 Nicole d' Oresme (1323–1382)
 Luca Pacioli (1445–1517)
 Blaise Pascal (1623–1662)
 Giuseppe Peano (1858–1932)
 J. Henri Poincaré (1854–1912)
 Jean-Victor Poncelet (1788–1867)
 Alfred Pringsheim (1850–1941)
 Diadochus Proclus (411–485)
 Pythagoras of Samos (580BC–520BC)
 Joseph Raabe (1801–1859)
 G. F. Bernhard Riemann (1826–1866)
 David Ruelle (1935–)
 Paolo Ruffini (1765–1822)
 Carle Runge (1856–1927)
 Bertrand Russell (1872–1970)
 Claude Shannon (1916–2001)
 Waclaw Sierpinski (1882–1969)
 Thomas Jan Stieltjes (1856–1894)
 James Stirling (1692–1770)
 Jean-Charles-François Sturm (1803–1855)
 Niccolò Fontana (1500–1557), called Tartaglia
 Brook Taylor (1685–1731)
 Thales of Miletus (624BC–546BC)
 Charles de la Vallée-Poussin (1866–1962)
 Pierre Verhulst (1804–1849)
 François Viète (1540–1603)
 Vito Volterra (1860–1940)
 John Wallis (1616–1703)
 Karl Weierstrass (1815–1897)
 Hermann Weyl (1885–1955)
 Alfred N. Whitehead (1861–1947)
 Oscar Zariski (1899–1986)
 Ernst Zermelo (1871–1951)
 Max Zorn (1906–1993)

There exist many web sites dedicated to the history of mathematics, we mention, e.g., <http://www-history.mcs.st-and.ac.uk/~history>.

B. Bibliographical Notes

We collect here a few suggestions for the readers interested in deepening some of the topics treated in this volume.

Concerning *numerical systems* the reader may consult

- H .E Ebbinghens, H. Hermes, F. Hirzebruch, M. Kaecher, K. Meier, J. Newrich, A. Prestel, R. Remmert, *Numbers*, Springer, New York, 1988.

We have discussed just a few elementary facts of the *theory of numbers*, for a thorough introduction we refer the reader for example to

- T. Apostol, *Introduction to Analytic Number Theory*, Springer, New York, 1976,
- G. H. Hardy and E. M. Wright, *An Introduction to the Theory of Numbers*, Oxford University Press, Oxford, 1938,
- A. Weyl, *Number Theory*, Birkhäuser, Boston, 1984.

The reader will easily find many books treating *combinatorics*, and, related to it, *discrete probability*; we mention a few titles

- W. Feller, *An Introduction to Probability Theory and its Applications*, John Wiley & Sons Inc., New York, 1957,
- C. L. Liu, *Introduction to Combinatorial Mathematics*, McGraw Hill, New York, 1968,
- J. Riordan, *An Introduction to Combinatorial Analysis*, John Wiley & Sons Inc., New York, 1958.

The reader will find surely quite stimulating

- R. Graham, D. Knuth and O. Patashnik, *Concrete Mathematics, a Foundation for Computer Science*, Addison-Wesley, Reading (Mass.), 1994,

that deals also with the *process of summation*.

A vast literature is available on *discrete dynamical processes* and *deterministic chaos*. We mention just a few titles. For a historical, cultural and popular approach the reader may consult

- A. Dahan Dalmedico, J. L. Chabert and K. Chemla Eds., *Chaos et déterminisme*, Éditions du Seuil, Paris, 1992,

and for a more technical approach

- W de Melo and S. von Shoen, *One-dimensional Dynamics*, Springer, Berlin, 1993,
- R. L. Devaney, *An Introduction to Chaotic Dynamical Systems*, The Benjamin Cummings Publishing Co., 1986.

C. Index

\aleph_0 , 101

Abel's

- test, 221
- theorem, 221, 224, 252, 253

algorithm

- Euclid, 4, 73, 310, 321
- for polynomials, 149
- Heron, 68
- logarithm-arcosin, 68
- Newton approximation, 314
- Pythagorean, 67
- quicksort, 313

almost-periodic orbit, 353

arc sine, 256

arc tangent, 197, 255

Archimedean property, 19

Argand's plane, 125

arrangements

- with repetitions, 90, 268
- without repetitions, 89, 268

asymptotic

- comparison test, 205
- ratio test, 230
- root test, 230

attractor, 355

- chaotic, 355

axiom

- continuity, 14, 15
- equivalent forms, 46
- continuum hypothesis, 106
- Dedekind's, 15
- of abstraction, 107
- of choice, 104, 108
- of extensionality, 107
- of infinity, 108
- of real numbers, 9
- of segregation, 108
- Peano's, 19
- Zermelo, 104

beating phenomenon, 134

Bell numbers, 271

Bernoulli's

- inequality, 22
- numbers, 235
- polynomials, 276
- shift, 351

Bessel's equation, 293

best rational approximation theorem, 326

beta function, 282

Bezout's theorem, 75

Binet's formula, 309

binomial

- coefficients, 33
- Newton's, 33
- series, 256
- theorem, 34

Bolzano-Weierstrass theorem, 46, 132

Cantor

- diagonal first method, 103
- diagonal second method, 105
- intersection theorem, 41
- middle-third set, 360
- principle, 41
- set, 230, 360, 365
- theorem, 105

Cantor-Bernstein theorem, 102

cardinal

- finite, 101
- transfinite, 101

cardinality, 88, 100

Catalan's identity, 29

Cauchy

- condensation test, 208
- criterion, 42
- sequence, 42

Cayley theorem, 118

Cesàro theorems, 51, 52, 67

characterization

- infimum, 14
- lower limit, 45
- supremum, 14

- upper limit, 44
- Chinese remainder theorem, 80
- combinations, 266
 - with replacements, 266
- combinatorics
 - drawings, 95
 - lists
 - increasing, 92
 - nondecreasing, 92
 - locations, 95
 - maps, 90
 - injective, 90
 - surjective, 94
 - nonordered k -samples
 - without replacement, 91
 - without replacements, 93
 - ordered k -samples
 - without repetitions, 89, 90
 - subsets, 91
- comparison test
 - for sequences, 37
 - for series, 204
- complex
 - differentiability, 152
 - hyperbolic functions, 259
 - trigonometric functions, 259
- complex numbers
 - n -th roots, 128
 - absolute value, 124
 - Argand plane, 125
 - beating phenomenon, 134
 - conjugate, 123
 - differential equations, 134
 - exponential, 127
 - Gauss plane, 122
 - Hermitian product, 125
 - imaginary unit, 122
 - modulus, 124
 - polar form, 125
 - product of, 126
 - prostapheresis, 134
 - roots of unity, 129
 - uniform circular motion, 133
- compound interest, 54
- congruences, 79
- continued fraction
 - convergent, 317
 - Euclid's algorithm, 321
 - periodic, 329
 - simple, 319
- continuum hypothesis, 106
- contraction
 - map, 338
 - mapping theorem, 338
- convergence
 - fast, 315
 - pointwise, 241
 - uniform, 241
- cosine, 254
- criterion
 - Cauchy, 42, 132
- curve
 - von Koch's, 230
- cut, 15
- d'Alembert lemma, 153
- de Moivre formula, 127
- Dedekind
 - axiom, 15
 - cut, 15
- dense subset, 20
- density
 - of decimal fractions, 21
 - of rationals, 20
- Descartes's law of signs, 164
- difference equation
 - first order, 302
 - second order, 304
- digamma, 285
- dimension
 - box counting, 357
 - Hausdorff, 359
- Dini–Riemann theorem, 226
- Dirichlet
 - kernel, 178
 - test, 221
 - theorem, 251
 - theorem about irrationals, 328
 - theorem on rearrangements, 225
 - theorem on series, 221
- discrete Jensen inequality, 28
- dispositions, 268
- distribution
 - hypergeometric, 97
 - multinomial, 99
- drawings, 95
- dynamical systems, 331
 - k -periodic
 - orbit, 341
 - point, 341
 - almost-periodic orbit, 353
 - attractor, 355
 - baker's map, 356
 - basin of attraction, 339
 - Bernoulli's shift, 351
 - chaotic
 - attractor, 355
 - orbit, 351
 - conjugate maps, 354
 - fixed point, 332
 - intermittency, 344
 - iterates, 332, 337
 - Lyapunov
 - exponent, 349

- number, 349
- orbit, 332
- sink, 339
- source, 339
- triangular map, 353, 356
- dynamics
 - contractive, 338
 - ergodic, 345
 - expansive, 338
- enumerator, 266
- equation
 - Bessel, 293
 - Euler, 293
 - fourth degree, 161
 - Laguerre, 293
 - Legendre, 292
 - second degree, 158
 - third degree, 159
- Eratosthenes sieve, 78
- ergodic theorem, 346
- Euclid
 - algorithm, 72, 310, 321
 - *for polynomials*, 149
 - generalized, 75
 - second theorem, 77
 - theorem, 73
- Euler's
 - Γ function, 280
 - equation, 293
 - formula, 127, 180
 - formula for cot, 213, 275
 - formula for sine, 212
 - function ϕ , 83
 - identity, 127
 - line, 137
 - numbers, 235
 - theorem, 84
- Euler-MacLaurin formula, 278
- exponential
 - complex, 127, 258
 - real, 60, 254
- factor theorem, 151
- factorial, 33
- Fagnano formula, 142
- Fatou theorem, 371
- Feigenbaum constant, 343
- Fermat minor theorem, 81
- Fibonacci numbers*, 308
- field
 - commutative, 123
 - ordered, 16
 - complete, 16
- fixed point, 331
- formula
 - Binet, 309
 - de Moivre, 127
 - Euler, 127, 180
 - for cot, 213
 - for sine, 212
 - on primes, 292
 - Euler-MacLaurin, 278
 - Fagnano, 142
 - Gauss, 281
 - Hermite, 168
 - inclusion-exclusion, 94
 - Legendre duplication, 284
 - of the complementary arguments, 284
 - Pascal, 33, 91
 - prostapheresis, 134
 - Simpson, 58
 - Stirling, 56, 280, 287
 - Vandermonde, 98
 - Viète, 211
 - Wallis, 55, 212
- fractals, 359
- Frobenius method, 293
- function
 - Γ , 280
 - ϕ of Euler, 82, 83
 - ζ of Riemann, 292
 - beta, 282
 - Dirichlet's, 174
 - generating, 264
 - harmonic, 174
 - periodic, 174
 - psi, 285
 - rational, 166
 - sinusoidal, 174
- fundamental sequence, 42
- fundamental theorem
 - of algebra, 156
 - of arithmetic, 77
- Γ function, 280
- gamma function
 - asymptotics, 287
- Gauss
 - formula, 281
 - plane, 122
 - psi function, 285
 - test, 234
- geometric progression, 54
- geometric series, 215
- graph, 116
 - *chromatic polynomial*, 117
 - coloring, 117
 - connected, 117
 - connected component, 117
- greatest common divisor, 72
 - for polynomials, 149
- group, 10
 - commutative, 10

- finite, 143
- generator, 143
- noncommutative, 11

Hardy theorem, 224

harmonic function, 174

Hausdorff

- dimension, 359

Hermite formula, 168

Hermitian product, 125

Heron's algorithm, 68

Hurwitz theorem, 328

ideal, 148

- principal, 149

identity

- Catalan, 29
- Euler's, 127
- Lagrange's, 28

induction principle, 18

inductive set, 17

inequality

- Bernoulli's, 22

infimum, 13

infinite product, 191

integral domain, 147

integration

- numerical
- rectangle rule, 57
- Simpson's rule, 58
- trapezoid formula, 57
- of rational functions, 171

intermediate value theorem, 49

isomorphism, 16

iterated function system, 361

- invariant set, 363

Jacobi's theorem, 346

Josephus problem, 24

Julia

- set, 371
- theorem, 371

k -repetitions, 89

k -samples

- nonordered
- without replacement, 91
- ordered, 89
- with replacement, 90
- without replacement, 89
- with replacements, 266
- without replacement, 266

Kronecker

- lemma, 232
- symbol, 177, 223

Lagrange's

- identity, 28
- interpolating polynomials, 186
- theorem about periodic continued fractions, 329

Laguerre equation, 293

Lamé theorem, 310

Legendre's

- duplication formula, 284
- equation, 292

Leibniz test, 216

limit, 35

- comparison test, 37
- constancy of sign, 37
- inferior, 44
- lower, 44
- of a sequence, 35, 36
- of monotone sequences, 39
- rules of calculus, 38
- squeezing test, 37
- subsequence, 41
- superior, 44
- uniform, 241
- continuity, 242
- uniqueness, 37
- upper, 44
- values, 45

Lindemann–Weierstrass theorem, 331

Liouville's theorem, 329

lists, 89

- increasing, 92

location

- distinct cells, 269

locations

- distinct cells, 95
- undistinct cells, 270

logarithm

- complex, 129, 259
- real, 63, 196, 255

logarithm-arccosin algorithm, 68

logistic model, 336

Lotka–Volterra models, 336

lower bound, 13

lower limit, 44

Lyapunov

- exponent, 349
- number, 349

Méré paradox, 114

Mandelbrot set, 372

maximum, 13

mean

- arithmetic, 22
- arithmetic-geometric, 68
- phase, 346
- quadratic, 22
- time, 346

Mertens theorem, 224

- minimum, 13
- Newton
 - approximation method, 314
 - binomial, 33
- Nicomachus theorem, 29
- nine-point circle theorem, 137
- nonlinear ODE
 - Euler method, 332
 - Heun method, 335
 - modified Euler method, 335
 - Runge–Kutta method, 335
- numbers
 - π , 260
 - e , 53, 62, 202, 260
 - algebraic, 103
 - bases 10, 193
 - Bell, 271
 - Bernoulli, 235
 - cardinal, 101
 - Carmichael's, 82
 - coprime, 72
 - decimal fractions, 21
 - decimals, 193
 - Euler's, 235
 - Euler–Mascheroni, 207
 - Feigenbaum, 343
 - Fibonacci, 308
 - integral, 20
 - irrationality of e , 203
 - natural, 17
 - prime, 72
 - pseudo-prime, 82
 - rational, 20
 - Stirling, 270
 - transcendental, 106
- orbit, 331
 - k -periodic, 341
 - chaotic, 351
- order
 - lexicographic, 123
 - partial, 104
 - total, 104
- ordered k -sample, 89
- ordered list, 92
- ovals, 28
- paradox
 - Méré, 114
 - Tarski, 28
- partitions, 271
 - of integers, 273
 - of sets, 271
- Pascal
 - formula, 33
 - triangle, 33
- Peano's axioms, 19
- permutation, 89
- phase mean, 346
- pointwise convergence, 241
- polynomials, 145
 - Bernoulli, 276
 - coercivity, 155
 - complex derivative, 152
 - coprime, 150
 - Euclid algorithm, 149
 - greatest common divisor, 149
 - Hermite, 293
 - irreducibility, 148
 - Lagrange's interpolating, 186
 - law of signs, 164
 - Legendre, 292
 - prime, 150
 - Sturm's sequence, 165
 - trigonometric, 175
 - energy equality, 177
 - sampling, 178
 - spectrum, 176
 - unique factorization, 150
- power of a set, 100
- power series
 - binomial, 256
 - boundary convergence test, 251
 - complex, 248
 - composition, 291
 - continuity of the sum, 243
 - derivative, 246
 - derivative of, 248
 - differential equations, 262
 - integral, 246
 - integral of, 248
 - inverse, 291
 - of derivatives, 244
 - of integrals, 244
 - radius of convergence, 238
 - reciprocal, 291
 - Taylor series, 247
 - uniform convergence, 241, 243
- powers
 - rational, 59
 - real, 60
- prime number theorem, 78
- principle
 - Cantor's, 41
 - exhaustion, 5
 - induction, 18
 - nested intervals, 41
 - of excluded middle, 108
 - of identity
 - of polynomials, 151
 - of power series, 248
- Pringsheim's theorem, 232
- product

- Cauchy, 222
- Hermitian, 125
- infinite, 191
- of convolution, 222
- property
 - Archimedean, 19
- psi function, 285
 - asymptotics, 287
- Pythagorean
 - algorithm, 67
 - theorem, 3
- quicksort algorithm, 313
- Raabe test, 234
- ratio test, 210
- reals
 - extended, 15
 - uniqueness, 16
- recursive
 - statements, 21
- relation
 - equivalence, 79, 100
 - order, 104
- root test, 209
- Roth's theorem, 331
- Ruffini theorem, 150
- semifactorials, 55
- sequence, 31
 - bounded, 39
 - above, 39
 - below, 39
 - bounded variation, 251
 - Cauchy, 42, 131
 - convergent, 35
 - decreasing, 39
 - divergent, 36
 - fundamental, 42
 - geometric, 50
 - increasing, 39
 - limit, 35
 - boundedness, 37
 - uniqueness, 37
 - lower limit, 44
 - maximizing, 40
 - minimizing, 40
 - monotone, 39
 - of complex numbers, 131
 - of partial sums, 189
 - product of convolution, 222
 - recursive, 32
 - strictly
 - decreasing, 39
 - increasing, 39
 - monotone, 39
 - subsequence, 41
 - total variation, 219
 - upper limit, 44
- series
 - Abel's test, 221
 - absolute convergence, 214, 216
 - alternating, 216
 - arithmetic-geometric, 191
 - asymptotic comparison test, 205
 - Cauchy condensation test, 208
 - comparison test, 204
 - convergent, 190
 - absolutely, 214, 216
 - decimal alignment, 193
 - Dirichlet test, 221
 - divergent, 190
 - domain of convergence, 240
 - Gauss's test, 234
 - generalized harmonic, 209
 - geometric, 190, 215, 254
 - harmonic, 207
 - improper integral, 192
 - indeterminate, 189
 - Leibniz test, 216
 - Mengoli, 190
 - nonnegative, 204
 - of complex terms, 215
 - partial sums, 189
 - Raabe test, 234
 - ratio test, 210
 - rearrangements, 225
 - root test, 209
 - sum, 189
 - summation by parts, 219
- set
 - bounded above, 13
 - bounded below, 13
 - Cantor, 230, 360, 365
 - cardinality, 100
 - chaotic, 355
 - countable, 101
 - dense, 20
 - denumerable, 101
 - equivalent, 100
 - finite, 101
 - inductive, 17
 - infimum, 13
 - infinite, 101
 - invariant, 363
 - Julia, 371
 - lower bound, 13
 - Mandelbrot, 372
 - maximum, 13
 - minimum, 13
 - partially ordered
 - chain, 105
 - maximal element, 105
 - maximum, 105

- supremum, 105
- upper bound, 105
- power of a , 100
- power of the continuum, 106
- self-similar, 361
- Sierpinski's
 - carpet, 362, 367
 - gasket, 362, 366
 - square, 362, 366
 - snowflake, 367
 - supremum, 13
 - upper bound, 13
 - von Koch's curve, 362, 368
- Sierpinski's
 - carpet, 367
 - gasket, 366
 - square, 366
- sieve of Eratosthenes, 78
- signal
 - amplitude, 175
 - amplitude spectrum, 175
 - fundamental harmonic, 175
 - harmonics, 175
 - phase spectrum, 175
 - pulse, 174
 - spectrum, 175, 176, 181
- sine
 - Euler's formula for sine, 212
 - real, 254
- sinusoidal signal, 174
 - amplitude, 174
 - phase, 174
 - pulse, 174
- statistics
 - Bose-Einstein, 96
 - Fermi-Dirac, 96
 - Maxwell-Boltzmann, 96
- Stirling
 - formula, 56, 280, 287
 - numbers, 270
- Sturm theorem, 165
- subsequence, 41
- sum
 - of a geometric progression, 54
 - of the arithmetic-geometric series, 191
 - of the first n naturals, 22
 - of the squares of the first n naturals, 24
- summation by parts, 219
- supremum, 13

- Tarski's paradox, 28
- Tartaglia triangle, 33
- Taylor series, 247
- Thale's theorem, 2
- theorem
 - Abel, 221, 224, 252, 253
 - best rational approximation, 326
 - Bezout, 75
 - binomial, 34
 - Bolzano-Weierstrass, 46, 132
 - boundary convergence, 251
 - Cantor, 105
 - Cantor's intersection, 41
 - Cantor-Bernstein, 102
 - Cauchy criterion, 42
 - Cayley, 118
 - Cesàro, 51, 52, 67
 - Chinese remainder, 80
 - contraction mapping, 338
 - d'Alembert, 153
 - differentiation term by term, 249
 - Dini-Riemann on rearrangements, 226
 - Dirichlet, 221
 - Dirichlet about irrationals, 328
 - Dirichlet on power series, 251
 - Dirichlet's on rearrangements, 225
 - ergodic, 346
 - Euclid, 73
 - second, 77
 - Euler, 84
 - factor, 151
 - Fatou, 371
 - Fermat minor, 81
 - fundamental of algebra, 156
 - fundamental of arithmetic, 77
 - Hardy, 224
 - Hurwitz, 328
 - integration term by term, 249
 - intermediate value, 49
 - Jacobi, 346
 - Julia, 371
 - Kronecker, 232
 - Lagrange about periodic continued fractions, 329
 - Lamé, 310
 - Lindemann-Weierstrass, 331
 - Liouville, 329
 - Mertens, 224
 - Nicomachus, 29
 - nine-point circle, 137
 - of exchanging limits and integrals, 245
 - prime number, 78
 - Pringsheim, 232
 - Pythagorean, 3
 - Roth, 331
 - Ruffini, 150
 - Sturm, 165
 - summation by parts, 219
 - term by term differentiation, 246
 - term by term integration, 246
 - Thales's, 2
 - two-squares, 143
 - Weierstrass, 49, 132
 - Weierstrass double series, 254

- well-ordering, 105
- Weyl, 346
- Zorn’s lemma, 105
- time mean, 346
- total variation, 219
- transform
 - \mathcal{Z} , 307
 - Laplace, 307
- tree, 118
- triangle
 - barycenter, 136
 - circumcenter, 136
 - Euler’s line, 137
 - Morley’s theorem, 139
 - Napoleon’s theorem, 138
 - nine-point circle, 137
 - orthocenter, 136
 - Pascal’s, 33, 91
 - Tartaglia, 33, 91
- triangular map, 353, 356
- trigonometric polynomial, 175
- two-squares theorem, 143

- uniform convergence, 241
- upper bound, 13
- upper limit, 44

- Vandermonde formula, 98
- Viète formula, 211
- von Koch’s curve, 230, 362

- Wallis’s formula, 55
- Weierstrass
 - double series theorem, 254
 - theorem, 49, 132
- Weyl theorem, 346

- Zorn’s lemma, 105

Mariano Giaquinta and Giuseppe Modica

Mathematical Analysis

Approximation and Discrete Processes

This fairly self-contained work embraces a broad range of topics in analysis at the graduate level, requiring only a sound knowledge of calculus and the functions of one variable. A key feature of this lively yet rigorous and systematic exposition is the historical accounts of ideas and methods pertaining to the relevant topics. Most interesting and useful are the connections developed between analysis and other mathematical disciplines, in this case, numerical analysis and probability theory.

The text is divided into two parts: The first examines the systems of real and complex numbers and deals with the notion of sequences in this context. After the presentation of natural numbers as a subset of the reals, elements of combinatorics and a discussion of the mathematical notion of the infinite are introduced. The second part is dedicated to discrete processes starting with a study of the processes of infinite summation both in the case of numerical series and of power series. The volume closes with an introductory chapter on the study of discrete dynamical systems and a summary of mathematicians and other scientists referenced in the work.

Mathematical Analysis: Approximation and Discrete Processes is replete with beautiful illustrations, examples, exercises at the end of each chapter, and a comprehensive index to aid the reader. It may be used in graduate seminars and courses or as a reference text by mathematicians, physicists, and engineers.

Birkhäuser

ISBN 0-8176-4313-3

www.birkhauser.com

