Applied Mathematics: Body & Soul
Vol I-III

K. Eriksson, D. Estep and C. Johnson

August 26, 2002

Preface

> I admit that each and every thing remains in its state until there is reason for change. (Leibniz)

The need of reform of mathematics education

Mathematics education needs to be reformed as we now pass into the new millenium. We share this conviction with a rapidly increasing number of researchers and teachers of mathematics and topics of science and engineering based on mathematical modeling. The reason is of course the computer revolution, which has fundamentally changed the possibilities of using mathematical and computational techniques for modeling, simulation and control of real phenomena. New products and systems may be developed and tested through computer simulation on time scales and at costs, which are orders of magnitude smaller than using traditional techniques based on extensive laboratory testing, hand calculations and trial and error.

At the heart of the new simulation techniques lie the new fields of Computational Mathematical Modeling CMM, including Computational Mechanics, Physics, Fluid Dynamics, Electromagnetics and Chemistry, all based on solving systems of differential equations using computers, combined with geometric modeling/Computer Aided Design CAD. Computational modeling is also finding revolutionary new applications in biology, medicine, environmental sciences, economy and finance markets.

The education in mathematics forms the basis of science and engineering education from undergraduate to graduate level, because engineering and science are largely based on mathematical modeling. The level and the quality of the mathematics education sets the level of the education as a whole. The new technology of CMM/CAD crosses borders between traditional engineering disciplines and schools, and puts strong forces to modernize engineering education in both content and form from basic to graduate level.

Our reform program

Our own reform work started some 20 years ago in courses in CMM on advanced undergraduate level, and has through the years successively penetrated through the system to the basic education in calculus and linear algebra. Our aim has become to develop a complete program for mathematics education in science and engineering from basic undergrauduate to graduate education. As of now our program contains the series of books:

1. *Computational Differential Equations*, (*CDE*)

2. *Applied Mathematics: Body* & *Soul I-III*, (*AM I-III*)

3. *Applied Mathematics: Body* & *Soul VI-*, (*AM IV-*).

AM I-III is the present book in three volumes I-III covering the basics of calculus and linear algebra. *AM IV-* offers a continuation with a series of volumes dedicated to specific areas of applications such as *Dynamical Systems (IV)*, *Fluid Mechanics (V)*, *Solid Mechanics (VI)* and *Electromagnetics (VII)*, which will start appearing in 2003. *CDE* published in 1996 may be be viewed as a first version of the whole *Applied Mathematics: Body* & *Soul* project.

Our program also contains a variety of software (collected in the *Mathematics Laboratory*), and by complementary material with step-by step instructions for self-study, problems with solutions, and projects, all freely available on-line. Our ambition is to be able to offer a "box" containing a set of books, software and additional instructional material, which can serve as a basis for a full applied mathematics program in science and engineering from basic to graduate level. Of course, we hope this to be an on-going project with new material being added gradually.

We have been running an applied mathematics program based on *AM* from first year for the students of chemical engineering at Chalmers since the Fall 99, and we have used collectedvarious advanced undergraduate courses

Main features of the program

- the program is based on a synthesis of mathematics, computation and application
- the program is based on new literature, giving a new unified presentation from start based on constructive mathematical methods including a computational methodology for differential equations
- the program contains, as an integrated part, software on different levels of complexity
- the student acquires solid skills of implementing computational methods and developing applications and software on a Matlab platform
- the synthesis of mathematics and computation opens the mathematics education to applications, and gives a basis for the effective use of modern mathematical methods in mechanics, physics, chemistry and applied subjects
- the synthesis building on constructive mathematics gives a synergetic effect allowing the study of complex systems already in the basic education, including the basic models of mechanical systems, heat conduction, wave propagation, elasticity, fluid flow, electro-magnetism, reaction-diffusion, molecular dynamcis and corresponding coupled multi-physics problems
- the program increases the motivation of the student by applying mathematical methods to interesting and important concrete problems already from the start
- emphasis may be put on problem solving, project work and presentation
- the program gives theoretical and computational tools and builds confidence
- the program contains most of the traditional material from basic courses in analysis and linear algebra
- the program includes much material often left out in traditional programs such as constructive proofs of all the basic theorems in analysis and linear algebra and advanced topics such as nonlinear systems of algebraic/differential equations
- emphasis is put on giving the student a solid understanding of basic mathematical concepts such as real numbers, Cauchy sequences, Lipschitz continuity, and constructive tools for solving algebraic/differential equations, together with an ability to utilize these tools in advanced applications such as molecular dynamics.

- the program may be run on different levels of ambition concerning both mathematical analysis and computation, while keeping a common basic core.

AM I-III in brief

Roughly speaking, *AM I-III* contains a synthesis of calculus and linear algebra including computational methods and a variety of applications. Emphasis is put on constructive/computational methods with the double aim of making the mathematics both understandable and useful. Our ambition is to introduce the student early to both (from the perspective of traditional education) advanced mathematical concepts (such as Lipschitz continuity, Cauchy sequence, contraction mapping, initial-value problem for systems of differential equations), and advanced applications such as Lagrangian mechanics, n-body systems, population models, elasticity and electrical circuits, with an approach based on constructive/computational methods.

The idea is thus that making the student comfortable with both advanced mathematical concepts and modern computational techniques, will open a wealth of possibilities of applying mathematics to problems of real interest. This is in contrast to the traditional education where the emphasis is usually put on a set of analytical techniques within a conceptual framework of more limited scope. For example: we lead already in the second quarter the student to write (in Matlab) his/her own solver for general systems of ordinary differential equations based on mathematically sound principles (high conceptual and computational level), while traditional education at the same time often focuses on training the student to master a bag of tricks for symbolic integration. We also teach the student some tricks to that purpose, but our over-all goal is different.

Constructive mathematics: Body & Soul

In our work we have been led to the conviction that the constructive aspects of calculus and linear algebra need to be strengthened. Of course, constructive and computational mathematics are closely related and the development of the computer has boosted computational mathematics in recent years. Mathematical modeling has two basic dual aspects: one symbolic and the other constructive-numerical, which reflect the duality between the infinite and the finite, or the continuum and the discrete. The two aspects have been closely intertwined throughout the development of modern science from the development of calculus in the work of Euler, Lagrange, Laplace and Gauss into the work of von Neumann in our time. For example, Laplace's monumental *Mécanique Céleste* in five volumes presents a symbolic calculus for a mathematical model of gravitation taking the form of Laplace's equation, together with massive numerical computations giving concrete information concerning the motion of the planets in our solar system.

However, beginning with the search for rigor in the foundations of calculus in the 19th century, a split between the symbolic and constructive aspects gradually developed. The split accelerated with the invention of the electronic computer in the 1940s, after which the constructive aspects were pursued in the new fields of numerical analysis and computing sciences, primarily developed outside departments of mathematics. The unfortunate result today is that symbolic mathematics and constructive-numerical mathematics by and large are separate disciplines and are rarely taught together. Typically, a student first meets calculus restricted to its symbolic form and then much later, in a different context, is confronted with the computational side. This state of affairs lacks a sound scientific motivation and causes severe difficulties in courses in physics, mechanics and applied sciences building on mathematical modeling.

Creating from start a synthesis of constructive and symbolic mathematics representing a synthesis of Body & Soul, opens new possibilities: with computational techniques available the students may become familiar with nonlinear systems of differential equations already in early calculus, with a wealth of applications. Another consequence is that the basics of calculus including concepts like real number, Cauchy sequence, convergence, fixed point iteration, contraction mapping, is lifted out of the wardrobe of mathematical obscurities into the real world with direct practical importance. In one shot one can make mathematics education both deeper, broader and lift it to a higher level. This idea is underlying the present book, which thus in the setting of a standard engineering program, contains all the basic theorems of calculus including the proofs normally taught only in special honours courses, together with advanced applications such as systems of nonlinear differential equations. We have found that this seemingly impossible program indeed works surprisingly well. Admittedly, this is hard to believe without making real life experiments. We hope the reader will feel encouraged to do so.

Lipschitz continuity and Cauchy sequences

The usual definition of the basic concepts of *continuity* and *derivative*, which is presented in most Calculus text books today, build on the concept of *limit*: a function of a real valued function $f(x)$ of a real variable x is said to be continuous at $\bar{x}$ if $\lim_{x\to\bar{x}} f(x) = f(\bar{x})$, and $f(x)$ is said to be differentiable at $\bar{x}$ with derivative $f'(\bar{x})$ if

$$\lim_{x \to \bar{x}} \frac{f(x) - f(\bar{x})}{x - \bar{x}}$$

exists and equals $f'(\bar{x})$. We use different definitions, where the concept of limit does not intervene: we say that $f(x)$ is Lipschitz continuous with Lipschitz constant L_f on the interval $[a, b]$ if for all $x, \bar{x} \in [a, b]$, we have

$$|f(x) - f(\bar{x})| \leq L_f |x - \bar{x}|.$$

Further, we say that $f(x)$ is differentiable at $\bar{x}$ with derivative $f'(\bar{x})$ if there is a constant $K_f(\bar{x})$ such that for all x close to $\bar{x}$

$$|f(x) - f(\bar{x}) - f'(\bar{x})(x - \bar{x})| \leq K_f(\bar{x})|x - \bar{x}|^2.$$

This means that we put somewhat more stringent requirements on the concepts of continuity and differentiability than is done in the usual definitions; more precisely, we impose *quantitative* measures in the form of the constants L_f and $K_f(\bar{x})$, whereas the usual definitions using limits are *purely qualitative.*

Using these more stringent definitions we avoid pathological situations, which can only be confusing to the student (in particular in the beginning), and as indicated, we avoid using the (difficult) concept of limit in a setting where in fact no limit processes are really taking place. Thus, we do not lead the student to definitions of continuity and differentiability suggesting that all the time the variable x is tending to some value $\bar{x}$, that is, that all the time some kind of (strange?) limit process is taking place. In fact, continuity expresses that the difference $f(x)-f(\bar{x})$ is small if $x-\bar{x}$ is small, and differentiability expresses that $f(x)$ locally is close to a linear function, and to express these facts we do not have to invoke any limit processes.

These are examples of our leading philosophy of giving Calculus a *quantitative form*, instead of the usual purely qualitative form, which we believe helps both understanding and precision. We believe the price to pay for these advantages is usually well worth paying, and the loss in generality are only some pathological cases of little interest. We can in a natural way relax our definitions for example to Hölder continuity, while still keeping the quantitative aspect, and thereby increase the pathology of the exceptional cases.

The usual definitions of continuity and differentiability strive for maximal generality, typically considered to be a virtue by a pure mathematician, which however has pathological side effects. With a constructive point of view the interesting world is the constructible world and maximality is not an important issue in itself.

Of course, we do not stay away from limit processes, but we concentrate on issues where the concept of limit really is central, most notably in defining the concept of a *real number* as the limit of a Cauchy sequence of rational numbers, and a solution of an algebraic or differential equation as the limit of a Cauchy sequence of approximate solutions. Thus, we give the concept of *Cauchy sequence* a central role, while maintaining a constructive approach always seeking a constructive process for generating a Cauchy sequence.

In standard Calculus texts, the concepts of Cauchy sequence and Lipschitz continuity are not used, believing them to be too difficult to be presented to freshmen, while the concept of real number is left undefined (seemingly believing that a freshman is so familiar with this concept from early life that no further discussion is needed). In contrast, in our constructive approach these concepts play a central role already from start, and in particular we spend a good deal of attention to the fundamental aspect of contructibilty of real numbers (viewed as possibly neverending decimal expansions).

We emphasize that taking a constructive approach does not make mathematical life more difficult in any important way, as is often claimed by the ruling mathematical school of formalists/logicists: All theorems of interest in Calculus and Linear Algebra survive, with possibly some small unessential modifications to keep the quantitative aspect and make the proofs more precise. As a result we are able to present basic theorems such as Contraction Mapping Principle, Implicit Function Theorem, Inverse Function Theorem, Convergence of Newton's Method, in a setting of several variables with complete proofs as a part of our basic Calculus, while these results in the standard curriculum are considered to be way to difficult for this level.

Proofs and Theorems

Most mathematics books including Calculus texts follow a theorem-proof style, where first a theorem is presented and then a corresponding proof is given. This is seldom appreciated very much by the students, who often have difficulties with the role and nature of the proof concept.

We usually turn this around and first present an line of thought leading to some result, and then we state a corresponding theorem as a summary of the hypothesis and the main result. We thus rather use a proof-theorem format. We believe this is in fact often more natural than the theorem-proof style, since by first presenting the line of thought the different ingredients like hypothesis may be introduced in a logical order. The proof will then be just like any other line of thought, where one successively derives consequences from some starting point using different hypothesis as one goes along. We hope this will help to eliminate the often perceived mystery of proofs, simply because the student will not be aware of the fact that that a proof is being presented; it will just a logical line of thought, like any logical line of thought in everyday life. Only when the line of thought is finished, one may go back and call it a proof, and in a theorem collect the main result arrived at including the required hypothesis. As a consequence, in the Latex version of the book we do use a theorem-environment, but not any proof-environment; the proof is just a logical line of thought preceeding a theorem collecting the hypothesis and the main result.

The Mathematics Laboratory

We have developed various pieces of software to support our program into what we refer to as the *Mathematics Laboratory.* Some of the software serves the purpose of illustrating mathematical concepts such as roots of equations, Lipschitz continuity, fixed point iteration, differentiability, the definition of the integral, basis calculus for functions of several varibles, other pieces are supposed to be used as models for the students own computer realizations, and finally some pieces are aimed at applications such as solvers for differential equations. New pieces are being added continuously. Our ambition is to also add different multi-media realizations of various parts of the material.

In our program the students get a training from start in using Matlab as a tool for computation. The development of the constructive mathematical aspects of the basic topics of real numbers, functions, equations, derivatives and integrals, goes hand in hand with experience of solving equations with fixed point iteration or Newton's method, quadrature, and numerical methods or differential equations. The students see from their own experience that abstract symbolic concepts have roots down into constructive computation, which also gives a direct coupling to applications and physical reality.

Acknowledgment

The authors of this book want to sincerely thank the following collegues and graduate students for contributing valuable material, corrections and suggestions for improvement: Rickard Bergström, Niklas Eriksson, Johan Hoffman, Mats Larson, Stig Larsson, Mårten Levenstam, Anders Logg, Klas Samuelsson and Nils Svanstedt, all actively participating in the development our reform project. And again, sincere thanks to all the students of chemical engineeering at Chalmers who carried the burden of being exposed to new material often in incomplete form, and who have given much of enthusiastic criticism and feed-back.

My heart is sad and lonely
for you I sigh, dear, only
Why haven't you seen it
I'm all for you body and soul
(Green, Body and Soul)

Contents

Preface **iii**

1 What is mathematics? **1**
1.1 Introduction . . . 1
1.2 The modern world: automatized production and computation 1
1.3 The role of mathematics . . . 4
1.4 Design and production of cars . . . 5
1.5 Navigation: from stars to GPS . . . 9
1.6 Medical tomography . . . 9
1.7 Molecular dynamics and medical drug design . . . 11
1.8 Weather prediction and global warming . . . 12
1.9 Economy: stocks and options . . . 12
1.10 Languages . . . 13
1.11 Mathematics as the language of science . . . 14
1.12 The basic areas of mathematics . . . 14
1.13 What is science? . . . 15
1.14 What is conscience? . . . 16
1.15 How to come to grips with the difficulties of understanding the material of this book and eventually view it as a good friend . . . 16

2 The Mathematics Laboratory **21**
2.1 Introduction . . . 21
2.2 Math Experience . . . 22

3 Introduction to modeling **27**
3.1 Introduction . . . 27
3.2 The Dinner Soup model . . . 27
3.3 The Muddy Yard model . . . 30
3.4 A system of equations: Dinner Soup/Ice Cream model . . . 31
3.5 Formulating and solving equations . . . 32

4 A very short Calculus course **35**
4.1 Introduction . . . 35
4.2 Algebraic equations . . . 36
4.3 Differential equations . . . 36
4.4 Generalization . . . 41
4.5 Leibniz' teen-age dream . . . 43
4.6 Summary . . . 45
4.7 Leibniz: Inventor of Calculus and Universal Genious . . . 46

5 Natural Numbers and Integers **49**
5.1 Introduction . . . 49
5.2 The natural numbers . . . 50
5.3 Is there a largest natural number? . . . 53
5.4 The set $\mathbb{N}$ of all natural numbers . . . 54
5.5 Integers . . . 55
5.6 Absolute value and the distance between numbers . . . 58
5.7 Division with remainder . . . 59
5.8 Factorization into prime factors . . . 60
5.9 Computer representation of integers . . . 61

6 Mathematical induction **65**
6.1 Induction . . . 65
6.2 Changes in a population of insects . . . 70

7 Rational numbers **73**
7.1 Introduction . . . 73
7.2 How to construct the rational numbers . . . 74
7.3 On the need for rational numbers . . . 77
7.4 Decimal expansions of rational numbers . . . 77
7.5 Periodic decimal expansions of rational numbers . . . 78
7.6 Set notation . . . 82
7.7 The set $\mathbb{Q}$ of all rational numbers . . . 83
7.8 The rational number line and intervals . . . 84
7.9 Growth of bacteria . . . 85
7.10 Chemical equilibrium . . . 87

8 Pythagoras and Euclid **89**
8.1 Introduction . . . 89

8.2 Pythagoras theorem 89
8.3 The sum of the angles of a triangle is 180° 91
8.4 Similar triangles 93
8.5 When are two straight lines orthogonal? 94
8.6 The GPS Navigator 96
8.7 Geometric definition of $\sin(v)$ and $\cos(v)$ 99
8.8 Geometric proof of addition formulas for $\cos(v)$ 99
8.9 Remembering some area formulas 101
8.10 Greek mathematics 101
8.11 The Euclidean plane $\mathbb{Q}^2$ 103
8.12 From Pythagoras to Euclid to Descartes 104
8.13 Non-Euclidean geometry 104

9 What is a Function? 107
9.1 Introduction 107
9.2 Functions in daily life 110
9.3 Graphing functions of integers 113
9.4 Graphing functions of rational numbers 117
9.5 A function of two variables 119
9.6 Functions of several variables 120

10 Polynomial functions 123
10.1 Introduction 123
10.2 Linear polynomials 124
10.3 Parallel lines 128
10.4 Orthogonal lines 129
10.5 Quadratic polynomials 130
10.6 Arithmetic with polynomials 134
10.7 Graphs of general polynomials 140
10.8 Piecewise polynomial functions 142

11 Combinations of functions 147
11.1 Introduction 147
11.2 Sum of two functions and product of a function with a number 148
11.3 Linear combinations of functions 148
11.4 Multiplication and division of functions 149
11.5 Rational functions 149
11.6 The composition of functions 152

12 Lipschitz continuity 155
12.1 Introduction 155
12.2 The Lipschitz continuity of a linear function 156
12.3 The definition of Lipschitz continuity 157
12.4 Monomials 161
12.5 Linear combinations of functions 163

12.6 Bounded functions 164
12.7 The product of functions 166
12.8 The quotient of functions 166
12.9 The composition of functions 167
12.10 Functions of two rational variables 168
12.11 Functions of several rational variables 169

13 Sequences and limits 173
13.1 A first encounter with sequences and limits 173
13.2 Socket wrench sets 175
13.3 J.P. Johansson's adjustable wrenches 177
13.4 The power of language: from infinitely many to one 177
13.5 The $\epsilon - N$ definition of a limit 178
13.6 A converging sequence has a unique limit 182
13.7 Lipschitz continuous functions and converging sequences . . 183
13.8 Generalization to functions of two variables 184
13.9 Computing limits 185
13.10 Computer representation of rational numbers 188
13.11 Sonya Kovalevskaya: The first woman on a chair in mathematics 189

14 The squareroot of two 195
14.1 Introduction 195
14.2 $\sqrt{2}$ is not a rational number! 197
14.3 Computing $\sqrt{2}$ by the Bisection algorithm 198
14.4 The Bisection algorithm converges! 199
14.5 First encounters with Cauchy sequences 202
14.6 Computing $\sqrt{2}$ by the Deca-section algorithm 203

15 Real numbers 207
15.1 Introduction 207
15.2 Adding and subtracting real numbers 209
15.3 Generalization to $f(x, \bar{x})$ with f Lipschitz 211
15.4 Multiplying and dividing real numbers 212
15.5 The absolute value 212
15.6 Comparing two real numbers 212
15.7 Summary of arithmetics with real numbers 213
15.8 Why $\sqrt{2}\sqrt{2}$ equals 2 213
15.9 A reflection on the nature of $\sqrt{2}$ 214
15.10 Cauchy sequences of real numbers. 215
15.11 Extension from $f : \mathbb{Q} \to \mathbb{Q}$ to $f : \mathbb{R} \to \mathbb{R}$ 216
15.12 Lipschitz continuity of extended functions 217
15.13 Graphing functions $f : \mathbb{R} \to \mathbb{R}$ 218
15.14 Extending a Lipschitz continuous function 218
15.15 Intervals of real numbers 220

15.16 What is $f(x)$ if x is irrational? 221
15.17 The usual definition of continuity vs Lipschitz continuity . . 223

16 The Bisection algorithm for $f(x) = 0$ **227**
16.1 Bisection 227
16.2 An example 229
16.3 Computational cost 232

17 Do mathematicians quarrel?* **235**
17.1 Introduction 235
17.2 The formalists 237
17.3 The logicists and set theory 238
17.4 The constructivists 241
17.5 The Peano axiom system for natural numbers 243
17.6 Real numbers 243
17.7 Cantor vs Kronecker 244
17.8 Deciding whether a number is rational or irrational 246
17.9 The set of all possible books 247
17.10 Recipes and good food 248
17.11 The “new math” in elementary education 249
17.12 The search for rigor in mathematics 249
17.13 A non-constructive proof 250
17.14 Summary 251

18 The function $y = x^r$ **255**
18.1 The function $\sqrt{x}$ 255
18.2 Computing with the function $\sqrt{x}$ 256
18.3 Is $\sqrt{x}$ Lipschitz continuous on $\mathbb{R}^+$? 256
18.4 The function x^r, where $r = \frac{p}{q}$ is rational 257
18.5 Computing with the function x^r 257
18.6 Generalizing the concept of Lipschitz continuity 257

19 Fixed points and contraction mapping **259**
19.1 Introduction 259
19.2 Contraction mappings 260
19.3 Rewriting $f(x) = 0$ as $x = g(x)$ 261
19.4 Card Sales Model 262
19.5 Private Economy Model 263
19.6 Fixed Point Iteration in the Card Sales model 264
19.7 A contraction mapping has a unique fixed point 269
19.8 Generalization to $g : [a, b] \to [a, b]$ 272
19.9 Linear convergence in Fixed Point Iteration 272
19.10 Quicker convergence 272
19.11 Quadratic convergence 275

20 Analytic geometry in $\mathbb{R}^2$ **281**
20.1 Introduction . . . 281
20.2 Descartes, Inventor of Analytic Geometry . . . 282
20.3 Descartes: Dualism of Body and Soul . . . 282
20.4 The Euclidean plane $\mathbb{R}^2$ . . . 283
20.5 Surveyors and Navigators . . . 285
20.6 A first glimpse of vectors . . . 286
20.7 Ordered pairs as points or vectors/arrows . . . 288
20.8 Vector addition . . . 289
20.9 Representing vector addition via the Parallelogram Law . . 289
20.10 Multiplication of a vector by a real number . . . 290
20.11 The norm of a vector . . . 291
20.12 Polar representation of a vector . . . 292
20.13 Standard basis vectors in $\mathbb{R}^2$ . . . 294
20.14 Scalar product . . . 294
20.15 Properties of the scalar product . . . 295
20.16 Geometric interpretation of the scalar product . . . 296
20.17 Orthogonality and scalar product . . . 297
20.18 Projection of a vector onto a vector . . . 298
20.19 Rotation by 90° . . . 301
20.20 Rotation by an arbitrary angle θ . . . 302
20.21 Rotation by θ again! . . . 303
20.22 Rotating a coordinate system . . . 303
20.23 Vector product in $\mathbb{R}^2$ . . . 304
20.24 The area of a triangle with a corner at the origin . . . 306
20.25 The area of a general triangle . . . 307
20.26 The area of a parallelogram spanned by two vectors . . . 308
20.27 Straight lines in $\mathbb{R}^2$ . . . 308
20.28 Projection of a point onto a line . . . 311
20.29 When are two lines parallel? . . . 311
20.30 A system of two linear equations in two unknowns . . . 312
20.31 The concepts of linear independence and basis . . . 314
20.32 The connection to calculus in one variable . . . 315
20.33 Linear mapping $f : \mathbb{R}^2 \to \mathbb{R}$ . . . 316
20.34 Linear mapping $f : \mathbb{R}^2 \to \mathbb{R}^2$ . . . 316
20.35 Linear mapping $f : \mathbb{R}^2 \to \mathbb{R}^2$ and linear system of equations 317
20.36 A first encounter with 2×2 and 2×1 matrices . . . 317
20.37 First applications of matrix notation . . . 319
20.38 Addition of matrices . . . 320
20.39 Multiplication of a matrix by a real number . . . 320
20.40 Multiplication of two matrices . . . 320
20.41 The transpose of a matrix . . . 322
20.42 The transpose of a 2-column vector . . . 322
20.43 The identity matrix . . . 323
20.44 The inverse of a matrix . . . 323

20.45 Rotation in matrix form again! 323
20.46 A mirror in matrix form 324
20.47 Change of basis again! 325
20.48 Queen Christina 326

21 Analytic geometry in $\mathbb{R}^3$ **331**
21.1 Introduction 331
21.2 Vector addition and multiplication by a scalar 333
21.3 Scalar product and norm 333
21.4 Projection of a vector onto a vector 334
21.5 The angle between two vectors in $\mathbb{R}^3$ 334
21.6 Vector product 335
21.7 Geometric interpretation of the vector product 337
21.8 Connection of vector products in $\mathbb{R}^2$ and $\mathbb{R}^3$ 338
21.9 Volume of parallelepiped spanned by three vectors 338
21.10 The triple product $a \cdot b \times c$ 339
21.11 An explicit formula for the volume $V(a_1, a_2, a_3)$ spanned by three vectors a_1, a_2 and a_3 340
21.12 Lines in $\mathbb{R}^3$ 341
21.13 Projection of a point onto a line 342
21.14 Planes in $\mathbb{R}^3$ 342
21.15 The intersection of a line and a plane 345
21.16 Two intersecting planes define a line 345
21.17 Projection of a point onto a plane 346
21.18 Distance from a point to a plane 347
21.19 Rotation around a given vector 348
21.20 Lines and planes through the origin are subspaces of $\mathbb{R}^3$ . . 349
21.21 Systems of 3 linear equations in 3 unknowns 349
21.22 Solving a 3×3-system by Gaussian elimination 350
21.23 3×3 matrices: sum, product and transpose 352
21.24 Two different ways of viewing a system of linear equations . 354
21.25 Non-singular 3×3 matrices 355
21.26 The inverse of a matrix 355
21.27 Different bases in $\mathbb{R}^3$ 356
21.28 Linearly independent set of vectors $\{a_1, a_2, a_3\}$ in $\mathbb{R}^3$ 356
21.29 Orthogonal matrices 357
21.30 Linear transformations versus matrices 357
21.31 The scalar product is invariant under orthogonal transformations 358
21.32 Looking ahead to functions $f : \mathbb{R}^3 \to \mathbb{R}^3$ 359
21.33 Unity of Body and Soul 359

22 Complex numbers **365**
22.1 Introduction 365
22.2 Addition and multiplication of complex numbers 366

22.3 The triangle inequality for complex numbers 367
22.4 Open domains in $\mathbb{C}$ 368
22.5 Polar representation of complex numbers 368
22.6 Geometrical interpretation of multiplication of complex numbers 369
22.7 Complex conjugation 369
22.8 Division of complex numbers 370
22.9 The Fundamental Theorem of Algebra 370
22.10 Roots 371
22.11 Solving a quadratic equation $w^2 + 2bw + c = 0$ 371
22.12 Gösta Mittag-Leffler 371

23 The derivative 375
23.1 Rates of change 375
23.2 Paying taxes 376
23.3 Hiking 379
23.4 Definition of the derivative 379
23.5 The derivative of a linear function is constant 382
23.6 The derivative of x^2 is $2x$ 383
23.7 The derivative of x^n is nx^{n-1} 384
23.8 The derivative of $\frac{1}{x}$ is $-\frac{1}{x^2}$ for $x \neq 0$ 385
23.9 The derivative $f'(\bar{x})$ as a function of $\bar{x}$ 385
23.10 Denoting the derivative of $f(x)$ by $Df(x)$ 386
23.11 Denoting the derivative of $f(x)$ by $\frac{df}{dx}$ 387
23.12 The derivative as a limit of difference quotients 387
23.13 How to compute a derivative? 389
23.14 Uniform differentiability on an interval 391
23.15 Bounded derivative implies Lipschitz continuity 392
23.16 Connecting to a slightly different definition of the derivative 394
23.17 Swedenborg 394

24 Differentiation rules 397
24.1 Introduction 397
24.2 The Linear Combination rule 398
24.3 The Product rule 399
24.4 The Chain rule 400
24.5 The Quotient rule 401
24.6 Derivatives of derivatives: $f^{(n)} = D^n f = \frac{d^n f}{dx^n}$ 402
24.7 One-sided derivatives 403
24.8 Quadratic approximation: Taylor's formula of order two . . 404
24.9 The derivative of an inverse function 407
24.10 Implicit differentiation 408
24.11 Partial derivatives 409
24.12 A sum up so far 411

25 Newton's method **413**
25.1 Introduction 413
25.2 Convergence of fixed point iteration 413
25.3 Newton's method 414
25.4 Newton's method converges quadratically 415
25.5 A geometric interpretation of Newton's method 416
25.6 What is the error in an approximate root? 417
25.7 Stopping criterion 419
25.8 Globally convergent Newton methods 420

26 Galileo, Newton, Hooke, Malthus and Fourier **423**
26.1 Introduction 423
26.2 Newton's Law of motion 424
26.3 Galileo's Law of motion 424
26.4 Hooke's Law 427
26.5 Newton's Law plus Hooke's Law 428
26.6 Fourier's Law for heat flow 429
26.7 Newton and rocket propulsion 430
26.8 Malthus and population growth 432
26.9 Einstein's Law of motion 433
26.10Summary 434

I Integrals
Geometry in $\mathbb{R}^n$
$u(x_N) - u(x_0) = \int_{x_0}^{x_N} u'(x)\,dx$

$\approx \sum_{j=1}^{N} u'(x_{j-1})(x_j - x_{j-1})$
$a \cdot b = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$ **437**

27 The Integral **439**
27.1 Primitive functions and integrals 439
27.2 Primitive function of $f(x) = x^m$ for $m = 0, 1, 2, ...$ 443
27.3 Primitive function of $f(x) = x^m$ for $m = -2, -3, ...$ 444
27.4 Primitive function of $f(x) = x^r$ for $r \neq -1$ and $x > 0$ 444
27.5 A quick overview of the progress so far 445
27.6 A "very quick proof" of the Fundamental Theorem 445
27.7 A "quick proof" of the Fundamental Theorem 447
27.8 A proof of the Fundamental Theorem of Calculus 448
27.9 Comments on the notation 455
27.10Alternative computational methods 455
27.11The bicyclist's speedometer 456
27.12Geometrical interpretation of the integral 456

27.13 The integral as a limit of Riemann sums 459
27.14 An Analog Integrator 459

28 Properties of the integral **463**
28.1 Introduction 463
28.2 Reversing the order of upper and lower limits 464
28.3 The whole is equal to the sum of the parts 464
28.4 Integrating piecewise Lipschitz continuous functions 465
28.5 Linearity 466
28.6 Monotonicity 467
28.7 The triangle inequality for integrals 467
28.8 Differentiation and integration are inverse operations 468
28.9 Change of variables or substitution 469
28.10 Integration by parts 471
28.11 The Mean Value Theorem 472
28.12 A function with non-negative derivative is non-decreasing . 474
28.13 A function with zero derivative is constant 474
28.14 Bounded derivative implies Lipschitz continuity (again) . . 475
28.15 Taylor's theorem 475
28.16 October 29, 1675 478
28.17 The hodometer 479

29 The logarithm $\log(x)$ **483**
29.1 The definition of $\log(x)$ 483
29.2 The importance of the logarithm 484
29.3 Properties: $\log(ab) = \log(a) + \log(b)$ and $\log(a^r) = r\log(a)$. 485

30 Numerical quadrature **489**
30.1 Computing integrals 489
30.2 The integral as a limit of Riemann sums: again 493
30.3 The midpoint rule 494
30.4 Adaptive quadrature 495

31 The exponential function $\exp(x) = e^x$ **501**
31.1 Introduction 501
31.2 Construction of the exponential function $\exp(x)$ for $x \geq 0$. 503
31.3 Construction of $\exp(-x)$ for $x \geq 0$ 508
31.4 The exponential $\exp(x)$ for $x \in \mathbb{R}$ 508
31.5 The basic property of the exponential: $\exp(a+b) = \exp(a)\exp(b)$ 508
31.6 The inverse of the exponential is the logarithm 510
31.7 The function a^x with $a > 0$ and $x \in \mathbb{R}$ 511

32 Trigonometric Functions **515**
32.1 The defining differential equation 515
32.2 Trigonometric formulas 519

32.3 The functions $\tan(x)$ and $\cot(x)$ and their derivatives . . . 520
32.4 Inverses of trigonometric functions 521
32.5 The functions $\sinh(x)$ and $\cosh(x)$ 523
32.6 The hanging chain 524
32.7 Comparing solutions of $u'' + k^2u(x) = 0$ and $u'' - k^2u(x) = 0$ 525

33 The functions $\exp(z)$, $\log(z)$, $\sin(z)$ and $\cos(z)$ for $z \in \mathbb{C}$ 527
33.1 Introduction 527
33.2 Definition of $\exp(z)$ for $z \in \mathbb{C}$ 527
33.3 Definition of $\sin(z)$ and $\cos(z)$ for $z \in \mathbb{C}$ 528
33.4 de Moivres formula 528
33.5 Definition of $\log(z)$ 529

34 Techniques of Integration 531
34.1 Introduction 531
34.2 Rational functions: the simple cases 532
34.3 Rational functions: partial fractions 533
34.4 A product of a polynomial function and a trigonometric or exponential function 538
34.5 Combination of trigonometric and root functions 538
34.6 Products of exponential and trigonometric functions 538
34.7 Products of polynomials and logarithms 539

35 Solving differential equations using the exponential 541
35.1 Introduction 541
35.2 Generalization to $u'(x) = \lambda(x)u(x) + f(x)$ 542
35.3 The equation $u''(x) - u(x) = 0$ 546
35.4 The differential equation $\sum_{k=0}^{n} a_kD^ku(x) = 0$ 547
35.5 The differential equation $\sum_{k=0}^{n} a_kD^ku(x) = f(x)$ 548
35.6 Euler's differential equation 549

36 Improper integrals 551
36.1 Introduction 551
36.2 Integrals over unbounded intervals 551
36.3 Integrals of unbounded functions 553

37 Series 557
37.1 Introduction 557
37.2 Definition of convergent infinite series 558
37.3 Positive series 559
37.4 Absolutely convergent series 561
37.5 Alternating series 562
37.6 The series $\sum_{i=1}^{\infty} \frac{1}{i}$ theoretically diverges! 563
37.7 Abel 564

38 Scalar autonomous initial value problems **567**
38.1 Introduction . . . 567
38.2 An analytical solution formula . . . 568
38.3 Construction of the solution . . . 571

39 Separable scalar initial value problems **575**
39.1 Introduction . . . 575
39.2 An analytical solution formula . . . 576
39.3 Volterra-Lotka's predator-prey model . . . 578
39.4 A generalization . . . 579

40 The general initial value problem **583**
40.1 Introduction . . . 583
40.2 Determinism and materialism . . . 585
40.3 Predictability and computability . . . 585
40.4 Construction of the solution . . . 587
40.5 Computational work . . . 588
40.6 Extension to second order initial value problems . . . 589
40.7 Numerical methods . . . 589

41 Calculus Tool Bag I **593**
41.1 Introduction . . . 593
41.2 Rational numbers . . . 593
41.3 Real numbers. Sequences and limits . . . 594
41.4 Polynomials and rational functions . . . 594
41.5 Lipschitz continuity . . . 595
41.6 Derivative . . . 595
41.7 Differentiation rules . . . 595
41.8 Solving $f(x) = 0$ with $f : \mathbb{R} \to \mathbb{R}$ . . . 596
41.9 Integrals . . . 596
41.10The logarithm . . . 598
41.11The exponential . . . 598
41.12The trigonometric functions . . . 599
41.13List of primitive functions . . . 601
41.14Series . . . 601
41.15Separable scalar initial value problems . . . 602

42 Analytic geometry in $\mathbb{R}^n$ **603**
42.1 Introduction and survey of basic objectives . . . 603
42.2 Body/Soul and Artificial Intelligence . . . 606
42.3 $\mathbb{R}^n$ as a vector space. Addition and scalar multiplication. . . 606
42.4 Scalar product, orthogonality . . . 607
42.5 Cauchy's inequality . . . 608
42.6 Vector space of linear combinations of a set of vectors . . . 609
42.7 The standard basis . . . 610

42.8 Linear independence 611
42.9 Elimination of linearly dependent elements gives a basis of a subspace 611
42.10 Constructing a basis for $S(a_1, ..., a_n)$ by reduction to column echelon form 612
42.11 Determining $R(A)$ by reduction to column echelon form . . 614
42.12 Determining $N(A)$ by reduction to row echelon form 616
42.13 Gaussian elimination 618
42.14 A basis for $\mathbb{R}^n$ contains n vectors 618
42.15 Coordinates in different bases 620
42.16 Linear functions $f : \mathbb{R}^n \to \mathbb{R}$ 621
42.17 Linear transformations $f : \mathbb{R}^n \to \mathbb{R}^m$ 621
42.18 Matrices 622
42.19 Matrix calculus 622
42.20 The transpose of a linear transformation 625
42.21 Matrix norms 625
42.22 Lipschitz constant of a linear transformation 626
42.23 Volume in $\mathbb{R}^n$. Determinant. Permutations 627
42.24 Definition of the volume $V(a_1, ...a_n)$ 628
42.25 The volume $V(a_1, a_2)$ in $\mathbb{R}^2$ 629
42.26 The volume $V(a_1, a_2, a_3)$ in $\mathbb{R}^3$ 629
42.27 The volume $V(a_1, a_2, a_3, a_4)$ in $\mathbb{R}^4$ 630
42.28 The volume $V(a_1, ..., a_n)$ in $\mathbb{R}^n$ 630
42.29 The determinant of a triangular matrix 631
42.30 Computing $\det A = V(a_1, ..., a_n)$ by reduction to column echelon form 631
42.31 The magic formula $\det AB = \det A \det B$ 631
42.32 Test of linear independence 632
42.33 Cramer's solution formula for non-singular $n \times n$ systems. . 633
42.34 Inverse matrix 635
42.35 Projection onto a subspace 635
42.36 An equivalent characterization of the projection 636
42.37 Orthogonal decomposition. Pythagoras theorem 637
42.38 Properties of projections 638
42.39 Orthogonalization: the Gram-Schmidt process 639
42.40 Orthogonal matrix 639
42.41 Invariance of the scalar product under orthogonal transformations 640
42.42 The QR-decomposition 640
42.43 The Fundamental Theorem of Linear Algebra 640
42.44 Change of basis: coordinates and matrix representations . . 642
42.45 Least squares methods 643

43 The Spectral Theorem **647**
43.1 Eigenvalues and eigenvectors 647

43.2 Basis of eigenvectors 649
43.3 The Spectral Theorem for symmetric matrices: easy version 650
43.4 Applying the Spectral Theorem to an IVP 651
43.5 The Spectral Theorem for symmetric matrices: general case 652
43.6 The norm of a symmetric matrix 654
43.7 Extension to non-symmetric real matrices 655

44 Solving linear algebraic systems 657
44.1 Introduction 657
44.2 Direct methods 658
44.3 Direct methods for special systems 664
44.4 Iterative methods 667
44.5 Estimating the error of the solution 678

45 Linear Algebra Tool Bag 687
45.1 Linear algebra in $\mathbb{R}^2$ 687
45.2 Linear algebra in $\mathbb{R}^3$ 688
45.3 Linear algebra in $\mathbb{R}^n$ 688
45.4 Linear transformations and matrices 689
45.5 Determinant and volume 690
45.6 Cramer's formula 691
45.7 Inverse 691
45.8 Projection 691
45.9 The Fundamental Theorem of Linear Algebra 691
45.10The QR-decomposition 692
45.11Change of basis 692
45.12Least squares method 692
45.13Eigenvalues and eigenvectors 692
45.14The Spectral Theorem 692

46 The matrix exponential $\exp(xA)$ **695**
46.1 Analytic computation of $\exp(xA)$ if A is diagonalizable . . . 696
46.2 Properties of $exp(Ax)$ 698
46.3 Duhamel's formula 698

47 Lagrange and the Principle of Least Action* 701
47.1 Introduction 701
47.2 Mass-spring system 703
47.3 Pendulum with fixed support 704
47.4 Pendulum with moving support 705
47.5 The Principle of Least Action 705
47.6 Conservation of the total energy 707
47.7 The double pendulum 707
47.8 The two-body problem 708
47.9 Stability of the motion of a pendulum 709

48 N-body systems* **711**
48.1 Introduction . . . 711
48.2 Masses and springs . . . 712
48.3 The N-body problem . . . 714
48.4 Masses, springs and dashpots: small displacements . . . 715
48.5 Adding dashpots . . . 716
48.6 The linear oscillator . . . 717
48.7 The damped linear oscillator . . . 718
48.8 Extensions . . . 719

49 The Crash model* **723**
49.1 Introduction . . . 723
49.2 The simplified growth model . . . 724
49.3 The simplified decay model . . . 726
49.4 The full model . . . 727

50 Electrical circuits* **731**
50.1 Introduction . . . 731
50.2 The components: inductors, resistors and capacitors . . . 732
50.3 Building circuits. Kirchhoff's laws . . . 733
50.4 Mutual induction . . . 734

51 String theory* **737**
51.1 Introduction . . . 737
51.2 A linear system . . . 738
51.3 A soft system . . . 739
51.4 A stiff system . . . 739
51.5 Phase plane analysis . . . 740

52 Piecewise linear approximation **743**
52.1 Introduction . . . 743
52.2 Interpolation on $[0,1]$ . . . 744
52.3 The vector space of linear functions on $[0,1]$ . . . 744
52.4 The vector space of piecewise linear continuous functions on an interval . . . 749
52.5 The L_2 projection into the space V_h of piecewise linear functions on I . . . 751

53 FEM for two-point boundary value problems **757**
53.1 Introduction . . . 757
53.2 Initial boundary-value problems . . . 760
53.3 Stationary boundary value problems . . . 761
53.4 The Finite Element Method . . . 761
53.5 The discrete system of equations . . . 764
53.6 Discretizing problems with different boundary conditions . . . 767

53.7 Error estimates and adaptive error control 770
53.8 Discretization of time-dependent reaction-diffusion-convection problems 774
53.9 Non-linear reaction-diffusion-convection problems 775

II Calculus in several dimensions

$\Delta u = \nabla \cdot \nabla u$

$\frac{\partial u}{\partial t} - \Delta u = f$

$\frac{\partial^2 u}{\partial t^2} - \Delta u = f$

$i\frac{\partial u}{\partial t} - \Delta u + Vu = 0$

$\frac{\partial u}{\partial t} + u \cdot \nabla u - \nu \Delta u = f$

$\nabla \times H = J, \quad \nabla \cdot B = 0, \quad B = \mu H.$ **778**

54 Vector-valued functions of several real variables 781
54.1 Introduction 781
54.2 Curves in $\mathbb{R}^n$ 782
54.3 Different parameterizations of a curve 783
54.4 Surfaces in $\mathbb{R}^n$, $n \geq 3$ 784
54.5 Lipschitz continuity 785
54.6 Differentiability: Jacobian, gradient and tangent 786
54.7 The Chain rule 790
54.8 The Mean Value theorem 791
54.9 Direction of steepest descent and the gradient 792
54.10A minimum point is a stationary point 794
54.11The method of Steepest Descent to find a minimum point . 794
54.12Directional derivative 795
54.13Higher order partial derivatives 795
54.14Taylor's theorem 797
54.15Contraction mapping theorem 798
54.16Solving $f(x) = 0$ with $f : \mathbb{R}^n \to \mathbb{R}^n$ 799
54.17The Inverse Function theorem 800
54.18The Implicit Function theorem 801
54.19Newton's method 802
54.20Differentiation under the integral sign 803

55 Level curves/surfaces and the gradient **807**
55.1 Level curves 807
55.2 Local existence of level curves 809
55.3 Level curves and the gradient 809
55.4 Level surfaces 811
55.5 Local existence of level surfaces 811
55.6 Level surfaces and the gradient 812

56 Linearization and stability of initial value problems **815**
56.1 Introduction 815
56.2 Stationary solutions 816
56.3 Linearization at a stationary solution 816
56.4 Stability analysis in the case $f'(\bar{u})$ is symmetric 817
56.5 Stability factors 818
56.6 Stability of time-dependent solutions 821
56.7 Sum up 821

57 Adaptive solvers of initial value prolems **823**
57.1 Introduction 823
57.2 The cG(1) method 824
57.3 Adaptive time step control for cG(1) 826
57.4 A posteriori error estimate for cG(1) a linear scalar IVP . . 826
57.5 A posteriori error estimate for cG(1) for a general IVP . . . 829
57.6 A posteriori error estimates for backward Euler for a general IVP 830
57.7 Stiff initial value problems 832
57.8 On explicit time-stepping for stiff problems 834

58 Lorenz and the Essence of Chaos* **841**
58.1 Introduction 841
58.2 The Lorenz system 842
58.3 The accuracy of the computations 845
58.4 Computability of the Lorenz system 846
58.5 The Lorenz Challenge 848

59 The Solar System* **851**
59.1 Introduction 851
59.2 Newton's Equation 853
59.3 Einstein's Equation 855
59.4 The Solar System as a system of ODEs 856
59.5 Predictability and Computability 858
59.6 Adaptive time-stepping 861
59.7 Limits of computability and predictability 862

60 Optimization **865**

60.1 Introduction 865
60.2 Sorting if Ω is finite 866
60.3 What if Ω is not finite? 867
60.4 Existence of minimum point 868
60.5 The derivative is zero at an interior minimum point 868
60.6 The role of the Hessian 871
60.7 Minimization algorithms: steepest descent 872
60.8 Existence of a minimum value and a minimum point 873
60.9 Existence of greatest lower bound to a set bounded below . 875
60.10 Analysis of the constructibility of minimum value and minimum point 876
60.11 A decreasing bounded sequence converges! 876

61 The divergence, rotation and Laplacian 879
61.1 Introduction 879
61.2 The gradient, divergence, rotation and Laplacian in $\mathbb{R}^2$ . . . 880
61.3 The Laplacian in $\mathbb{R}^2$ in polar coordinates 881
61.4 Basic examples with $\nabla \cdot u = 1$, $\nabla \times v = 1$ and $\Delta w = 1$ . . . 882
61.5 Invariance of the Laplacian under rigid coordinate transformations 882
61.6 The gradient, divergence, rotation and Laplacian in $\mathbb{R}^3$ . . . 883
61.7 Basic examples with $\nabla \cdot u = 1$, $\nabla \times v = (0, 0, 1)$ and $\Delta w = 1$ 884
61.8 The Laplacian in spherical coordinates 884

62 Meteorology and Coriolis forces* 887
62.1 Introduction 887
62.2 A basic meteorological model 888
62.3 Rotating coordinate systems and Coriolis acceleration . . . 888

63 Curve integrals 893
63.1 Introduction 893
63.2 The length of a curve in $\mathbb{R}^2$ 893
63.3 Curve integral 895
63.4 Reparametrization 896
63.5 Work and gradient fields 898
63.6 Using the arclength as parameter 899
63.7 Extension to curves in $\mathbb{R}^n$ 901

64 Double integrals 905
64.1 Introduction 905
64.2 Double integrals over the unit square 906
64.3 The double integral over Q through repeated one-dimensional integration 909
64.4 Generalization to an arbitrary rectangle 911
64.5 Interpreting the double integral as a volume 911

64.6 Extension to general domains 912
64.7 Iterated integrals over general domains 914
64.8 The area of a two-dimensional domain 915
64.9 The integral as the limit of a general Riemann sum 916
64.10 Change of variables in a double integral 917

65 Surface integrals 923
65.1 Introduction 923
65.2 Surface area 923
65.3 The surface area of a function of two variables 925
65.4 Surfaces of revolution 926
65.5 Independence of parameterization 927
65.6 Surface integrals 927
65.7 Moment of inertia of a thin spherical shell 928

66 Multiple integrals 931
66.1 Introduction 931
66.2 Triple integrals over the unit cube 931
66.3 Triple integrals over more general domains in $\mathbb{R}^3$ 932
66.4 The volume of a three-dimensional domain 933
66.5 Triple integrals as limits of Riemann sums 934
66.6 Change of variables in a triple integral 935
66.7 Solids of revolution 937
66.8 Moment of inertia of a ball 938

67 Gauss' theorem and Green's formula in $\mathbb{R}^2$ 941
67.1 Introduction 941
67.2 The special case with Ω a square 942
67.3 The general case 942

68 Gauss' theorem and Green's formula in $\mathbb{R}^3$ 949
68.1 Introduction 949

69 Stokes' theorem 955
69.1 Introduction 955
69.2 The special case of a surface in a plane 956
69.3 Generalization to an arbitrary plane surface 958
69.4 Generalization to a surface bounded by a plane curve 958

70 Potential fields 963
70.1 Introduction 963
70.2 An irrotational field is a potential field 964
70.3 A counter-example for a non-convex Ω 966

71 Center of mass and Archimedes' principle* 967
71.1 Introduction 967

71.2 Center of mass 968
71.3 Archimedes' principle 971
71.4 Stability of floating bodies 973

72 Newton's nightmare* **977**

73 Laplacian models **983**
73.1 Introduction 983
73.2 Heat conduction 983
73.3 The heat equation 986
73.4 Stationary heat conduction: Poisson's equation 987
73.5 Convection-diffusion-reaction 989
73.6 Elastic membrane 990
73.7 Solving the Poisson equation 991
73.8 The wave equation: vibrating elastic membrane 993
73.9 Fluid mechanics 993
73.10 Maxwell's equations 999
73.11 Gravitation 1003
73.12 Quantum Mechanics 1007

74 Chemical reactions* **1013**
74.1 Constant temperature 1013
74.2 Variable temperature 1016
74.3 Space dependence 1017

75 Calculus Tool Bag II **1019**
75.1 Introduction 1019
75.2 Lipshitz continuity 1020
75.3 Differentiability 1020
75.4 The Chain rule 1020
75.5 Mean Value theorem for $f : \mathbb{R}^n \to \mathbb{R}$ 1020
75.6 A minimum point is a stationary point 1020
75.7 Taylor's theorem 1021
75.8 Contraction mapping theorem 1021
75.9 Inverse Function theorem 1021
75.10 Implicit Function theorem 1021
75.11 Newton's method 1021
75.12 Differential operators 1022
75.13 Curve integrals 1023
75.14 Multiple integrals 1023
75.15 Surface integrals 1024
75.16 Green's and Gauss' formulas 1024
75.17 Stokes' theorem 1025

76 Piecewise linears in $\mathbb{R}^2$ and $\mathbb{R}^3$ **1027**

76.1 Introduction 1027
76.2 Triangulation of a domain in $\mathbb{R}^2$ 1028
76.3 Mesh generation in $\mathbb{R}^3$ 1031
76.4 Piecewise linear functions 1031
76.5 Max-norm error estimates for piecewise linear interpolation 1034
76.6 Quadrature in $\mathbb{R}^2$ 1037

77 FEM for boundary value problems in $\mathbb{R}^2$ and $\mathbb{R}^3$ 1041
77.1 Introduction 1041
77.2 Variational formulation 1042
77.3 The cG(1) FEM 1043
77.4 Basic data structures 1049
77.5 Solving the discrete system 1049
77.6 An equivalent minimization problem 1051
77.7 An energy norm a priori error estimate 1051
77.8 An energy norm a posteriori error estimate 1052
77.9 Adaptive error control 1054
77.10An example 1056
77.11Extension to non-homogeneous Dirichlet boundary conditions1057
77.12The L-shaped membrane 1058
77.13Robin and Neumann boundary conditions 1059
77.14Stationary convection-diffusion-reaction problems 1061
77.15Time-dependent convection-diffusion-reaction problems . . . 1062
77.16The wave equation 1063
77.17Examples 1063

78 Inverse problems 1067
78.1 Introduction 1067
78.2 An inverse source problem for one-dimensional convection . 1069
78.3 An inverse source problem for one-dimensional diffusion . . 1071
78.4 An inverse source problem for Poisson's equation 1073
78.5 An inverse boundary value problem for Laplace's equation . 1075
78.6 The Backward heat equation 1076
78.7 Further applications 1077

79 Differential Equations Tool Bag 1079
79.1 Introduction 1079
79.2 The equation $u'(x) = \lambda(x)u(x)$ 1080
79.3 The equation $u'(x) = \lambda(x)u(x) + f(x)$ 1080
79.4 The differential equation $\sum_{k=0}^{n} a_k D^k u(x) = 0$ 1080
79.5 The damped linear oscillator 1081
79.6 The matrix exponential 1081
79.7 Fundamental solutions of the Laplacian 1082
79.8 The wave equation in 1d 1082
79.9 Numerical methods for initial value problems 1082

79.10 cg(1) for convection-diffusion-reaction 1083
79.11 Svensson's Formula for Laplace's equation 1083

80 Applications Tool Bag **1085**

80.1 Introduction 1085
80.2 Malthus' population model: 1085
80.3 The logistics equation 1085
80.4 Mass-spring-dashpot system 1085
80.5 LCR-circuit 1086
80.6 Laplace's equation for gravitation: 1086
80.7 The heat equation: 1086
80.8 The wave equation: 1086
80.9 Convection-diffusion-reaction: 1086
80.10 Maxwell's equations: 1087
80.11 The incompressible Navier-Stokes equations: 1087

81 Analytic functions **1089**

81.1 Introduction 1089
81.2 The definition of an analytic function 1089
81.3 The derivative as a limit of difference quotients 1091
81.4 The function $f(z) = az + b$ is analytic in $\mathbb{C}$ 1091
81.5 The function $f(z) = z^2$ is analytic in $\mathbb{C}$ 1091
81.6 The function $f(z) = z^n$ is analytic in $\mathbb{C}$ for $n = 1, 2, ...$ 1092
81.7 Rules of differentiation 1092
81.8 The function $f(z) = z^{-n}$ is analytic for $z \neq 0$ if $n = 1, 2, ...$ 1092
81.9 The Cauchy-Riemann equations 1092
81.10 The Cauchy-Riemann equations and the limit definition of the derivative 1094
81.11 The Cauchy-Riemann equations in polar coordinates 1095
81.12 The real and imaginary parts of an analytic function are harmonic 1095
81.13 The derivative of an analytic function is an analytic function 1096
81.14 Conjugate harmonic functions 1096
81.15 Curves in the complex plane 1097
81.16 Conformal mappings 1097
81.17 Translation-rotation-expansion/contraction 1099
81.18 Inversion 1099
81.19 Möbius transformations 1100
81.20 The functions $w = z^{1/2}$, $w = e^z$, $w = \log(z)$ and $w = \sin(z)$ 1101
81.21 Complex integrals: first shot 1102
81.22 Complex integrals: general case 1104
81.23 Basic properties of the complex integral 1105
81.24 Taylor's formula: first shot 1105
81.25 Cauchy's theorem 1106
81.26 Cauchy's representation formula 1107

81.27 Taylor's formula: second shot . . . 1108
81.28 Power series representation of analytic functions . . . 1110
81.29 Laurent series . . . 1111
81.30 Residue Calculus: simple poles . . . 1112
81.31 Residue Calculus: poles of any order . . . 1114
81.32 The Residue Theorem . . . 1114
81.33 Computation of $\int_0^{2\pi} R(\sin(t), \cos(t))\, dt$ . . . 1115
81.34 Computation of $\int_{-\infty}^{\infty} \frac{p(x)}{q(x)}\, dx$ . . . 1116
81.35 Applications to potential theory in $\mathbb{R}^2$ . . . 1117

82 Fourier series **1123**
82.1 Introduction . . . 1123
82.2 Warm up I: orthonormal basis in $\mathbb{C}^n$ . . . 1125
82.3 Warm up II: Series . . . 1126
82.4 Complex Fourier series . . . 1126
82.5 Fourier series as an orthonormal basis expansion . . . 1128
82.6 Truncated Fourier series and best L_2-approximation . . . 1128
82.7 Real Fourier series: Cosine and Sine series representations . 1129
82.8 Basic properties of Fourier coefficients . . . 1131
82.9 The Fourier coefficients $c_m(f)$ tend to zero as $|m| \to \infty$ . . 1133
82.10 Convolution . . . 1134
82.11 The inversion formula . . . 1135
82.12 Parsevals formula . . . 1137
82.13 Space vs frequency analysis . . . 1138
82.14 Fourier series of periodic functions with period different from 2π . . . 1139
82.15 Weierstrass functions . . . 1139
82.16 Solving the 1d heat equation using Fourier series . . . 1140
82.17 Computing Fourier coefficients with quadrature . . . 1141
82.18 The Discrete Fourier Transforms . . . 1142

83 Fourier transforms **1145**
83.1 Basic properties of the Fourier transform . . . 1147
83.2 The Fourier transform $\widehat{f}(\xi)$ tends to zero as $|\xi| \to \infty$ . . . 1148
83.3 The inversion formula . . . 1149
83.4 Parsevals formula . . . 1150
83.5 Solving the 1d heat equation on $\mathbb{R}$ using the Fourier transform 1151
83.6 The connection between Fourier series and Fourier transforms 1151
83.7 The sampling theorem . . . 1152
83.8 The Laplace transform . . . 1153
83.9 Wavelets and the Haar basis . . . 1154

84 Analytic Functions Tool Bag **1157**
84.1 Differentiability and analyticity . . . 1157
84.2 The Cauchy-Riemann equations . . . 1157

84.3 The real and imaginary parts of an analytic function are harmonic . . . 1158
84.4 Conjugate harmonic functions . . . 1158
84.5 Curves in the complex plane . . . 1158
84.6 An analytic function defines a conformal mapping . . . 1159
84.7 Complex integrals . . . 1159
84.8 Cauchy's theorem . . . 1159
84.9 Cauchy's representation formula . . . 1159
84.10 Taylor's formula . . . 1160
84.11 The Residue Theorem . . . 1160

85 Fourier Analysis Tool Bag 1161
85.1 Properties of Fourier coefficients . . . 1161
85.2 Convolution . . . 1161
85.3 Fourier series representation . . . 1162
85.4 Parseval's formula . . . 1162
85.5 Discrete Fourier Transforms . . . 1162
85.6 Fourier transforms . . . 1162
85.7 Properties of Fourier transforms . . . 1163
85.8 The Sampling Theorem . . . 1163

86 Incompressible Navier-Stokes: quick and easy 1165
86.1 Introduction . . . 1165
86.2 The incompressible Navier-Stokes equations . . . 1166
86.3 Numerical methods for the incompressible Navier-Stokes equations . . . 1167
86.4 The cG(1)dG(0) method . . . 1168
86.5 The cG(1)cG(1) method . . . 1169
86.6 The cG(1)dG(1) method . . . 1170
86.7 Neumann boundary conditions . . . 1170
86.8 Periodic boundary conditions . . . 1172
86.9 Computational examples . . . 1173

Index 1177

1
What is mathematics?

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 whether a final objective answer may be expected at all. "Mathematizing" may well be a creative activity of man, like language or music, of primary originality, whose historical decisions defy complete objective rationalization. (Weyl)

1.1 Introduction

We start out by giving a very brief idea of the nature of mathematics and the role of mathematics in our society.

1.2 The modern world: automatized production and computation

The mass consumption of the *industrial society* is made possible by the *automatized mass production* of material goods such as food, clothes, housing, TV-sets, CD-players and cars. If these items had to be produced by hand, they would be the privileges of only a select few.

Analogously, the emerging *information society* is based on mass consumption of *automatized computation* by computers that is creating a new "virtual reality" and is revolutionizing technology, communication, admin-

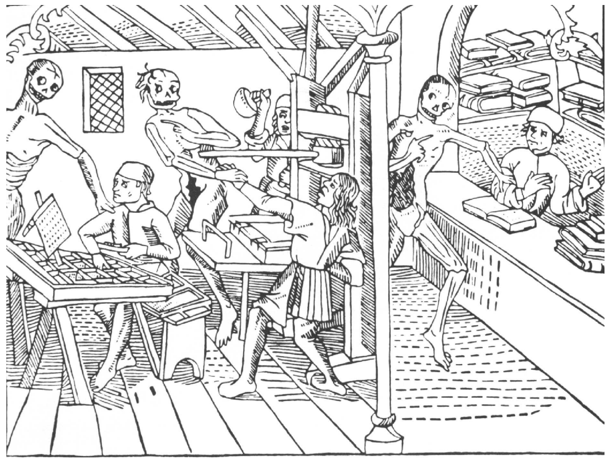


FIGURE 1.1. First picture of book printing technique (from Danse macabre, Lyon 1499)

istration, economy, medicine, and the entertainment industry. The information society offers immaterial goods in the form of knowledge, information, fiction, movies, music, games and means of communication. The modern PC or lap-top is a powerful computing device for mass production/consumption of information e.g. in the form of words, images, movies and music.

Key steps in the automatization or mechanization of production were: Gutenbergs's book printing technique (Germany, 1450), Christoffer Polhem's automatic machine for clock gears (Sweden, 1700), The Spinnning Jenny (England, 1764), Jacquard's punched card controlled weavingloom (France, 1801), Ford's production line (USA, 1913), see Fig. 1.1, Fig. 1.2, and Fig. 1.3.

Key steps in the automatization of computation were: Abacus (Ancient Greece, Roman Empire), Slide Rule (England, 1620), Pascals Mechanical Calculator (France, 1650), Babbage's Difference Machine (England, 1830), Scheutz' Difference Machine (Sweden, 1850), ENIAC Electronic Numerical Integrator and Computer (USA, 1945), and the Personal Computer PC (USA, 1980), see Fig. 1.5, Fig. 1.6, Fig. 1.7 and Fig. 1.8. The Difference Machines could solve simple differential equations and was used to compute tables of elementary functions such as the logarithm. ENIAC was one of the first modern computers (electronic and programmable), consisted of 18.000 vacuum tubes filling a room of 50×100 squarefeet with a weight of 30 tons and energy consuming of 200 kilowatts, and was used to solve the differential equations of ballistic firing tables as an important part of the Allied World War II effort.

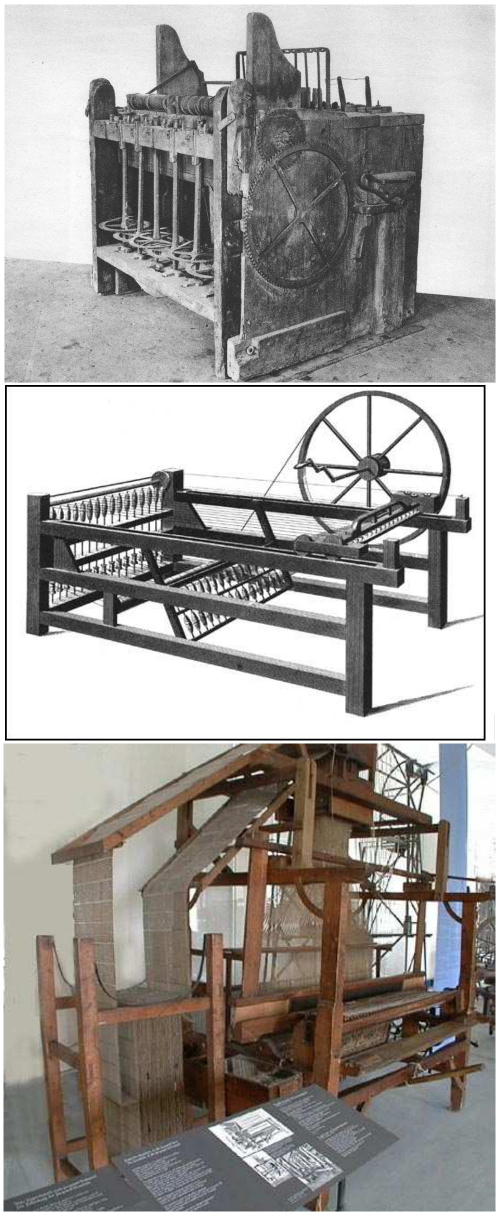


FIGURE 1.2. Christoffer Polhem's machine for clock gears (1700), Spinning Jenny (1764) and Jaquard's programmable loom (1801).

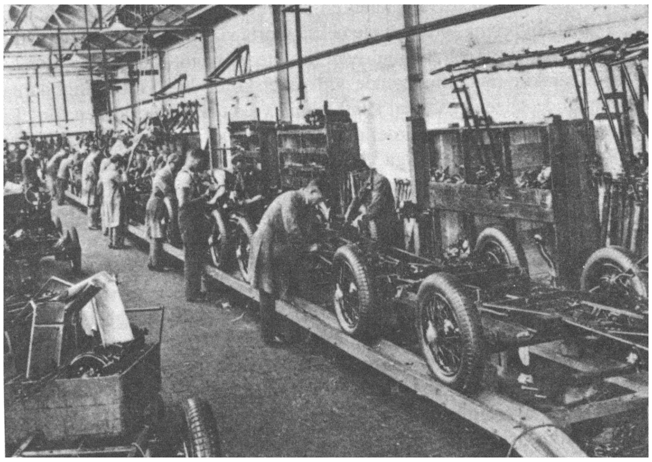


FIGURE 1.3. Ford assembly line (1913).

Automatization (or automation) is based on frequent repetion of a certain *algorithm* or scheme with new data at each repetition. The algorithm may consist of a sequence of relatively simple steps together creating a more complicated process. In automatized manufacturing, as in the production line of a car factory, physical material is modified following a strict repetitive scheme, and in automatized computation, the 1s and 0s of the microprocessor are modified billions of times each second following the computer program. Similarly, a *genetic code* of an organism may be seen as an algorithm that generates a living organism when realized in interplay with the environment. Realizing a genetic code many times (with small variations) generates populations of organisms. Mass-production is the key to increased complexity following the patterns of nature: elementary particle $\rightarrow$ atom $\rightarrow$ molecule and molecule $\rightarrow$ cell $\rightarrow$ organism $\rightarrow$ population, or the patterns of our society: individual $\rightarrow$ group $\rightarrow$ society or computer $\rightarrow$ computer network $\rightarrow$ global net.

1.3 The role of mathematics

Mathematics may be viewed as the language of computation and thus lies at the heart of the modern information society. Mathematics is also the language of science and thus lies at the heart of the industrial society that grew out of the *scientific revolution* in the 17th century that began when Leibniz and Newton created *Calculus*. Using Calculus, basic laws of mechanics and physics, such as Newton's law, could be formulated as *mathematical models* in the form of *differential equations*. Using the models, real phenomena

FIGURE 1.4. Computing device of the Inka Culture

could be *simulated* and controlled (more or less) and industrial processes could be created.

The mass consumption of both material and immaterial goods, considered to be a corner-stone of our modern democratic society, is made possible through automatization of production and computation. Therefore, mathematics forms a fundamental part of the technical basis of the modern society revolving around automatized production of material goods and automatized computation of information.

The vision of virtual reality based on automatized computation was formulated by Leibniz already in the 17th century and was developed further by Babbage with his Analytical Engine in the 1830s. This vision is finally being realized in the modern computer age in a synthesis of Body & Soul of Mathematics.

We now give some examples of the use of mathematics today that are connected to different forms of automatized computation.

1.4 Design and production of cars

In the car industry, a model of a component or complete car can be made using Computer Aided Design CAD. The CAD-model describes the ge-

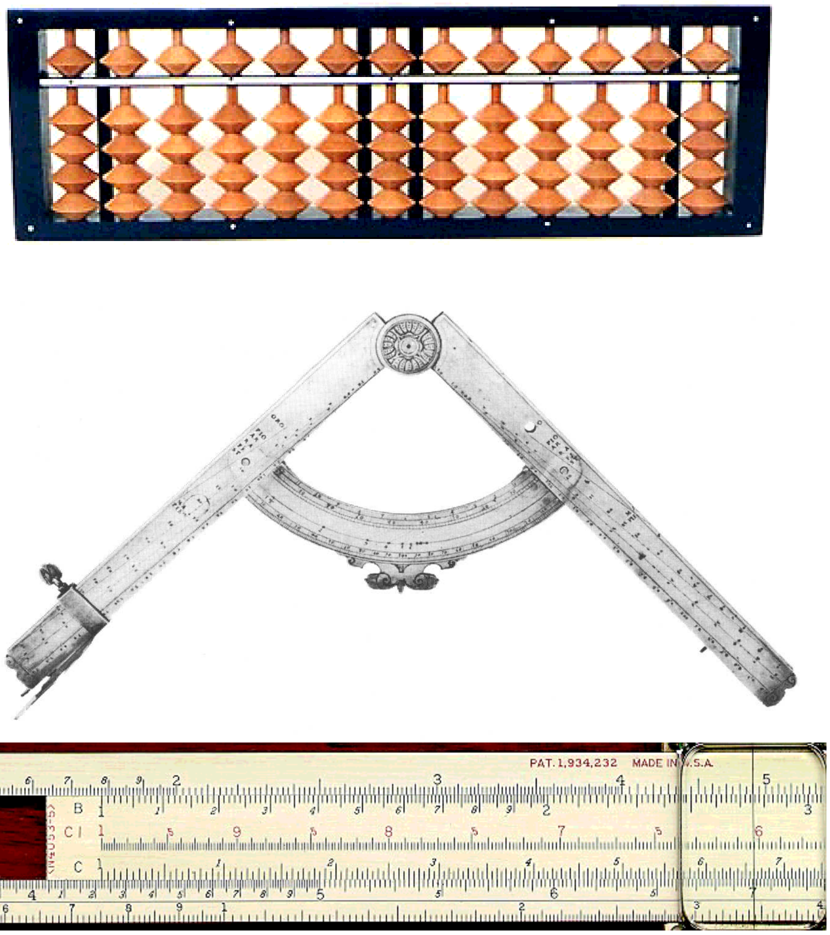


FIGURE 1.5. Classical computational tools: Abacus (300 B.C.-), Galileo's Compass (1597) and Slide Rule (1620-).

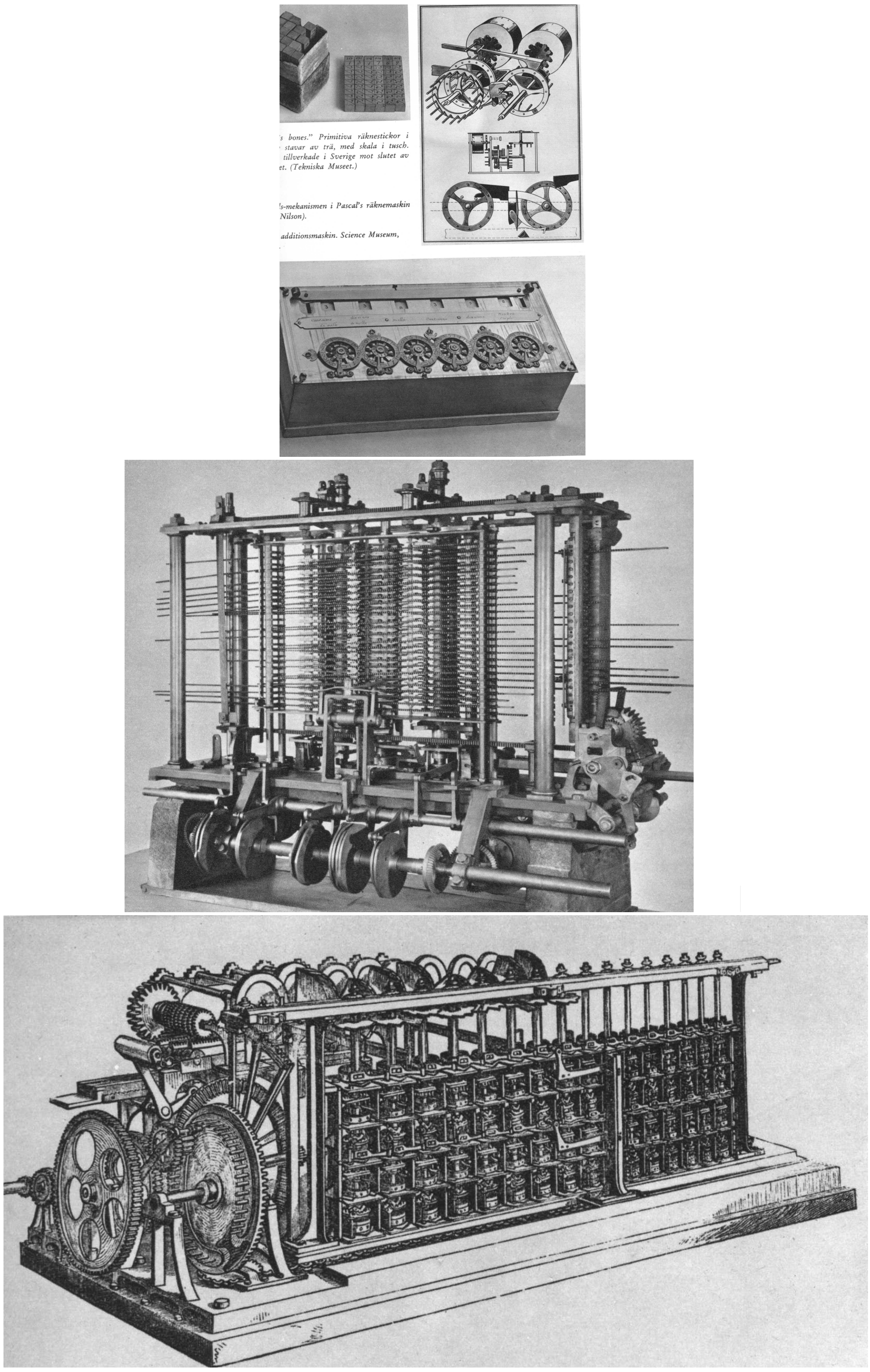


FIGURE 1.6. Napier's Bones (1617), Pascals Calculator (1630), Babbage's Difference Machine (1830) and Scheutz' Swedish Difference Machine (1850).

FIGURE 1.7. Odhner's mechanical calculator made in Göteborg, Sweden, 1919-1950

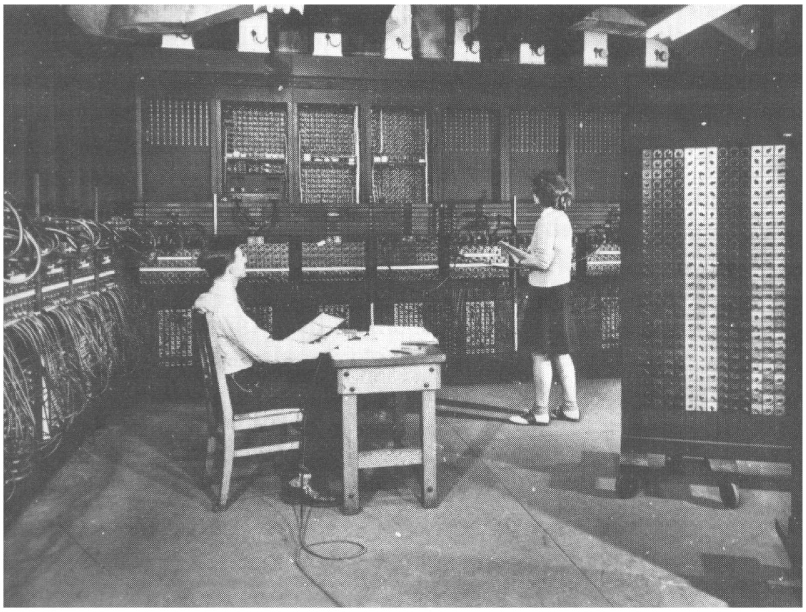


FIGURE 1.8. ENIAC Electronic Numerical Integrator and Calculator (1945)

ometry of the car through mathematical expressions and the model can be displayed on the computer screen. The performance of the component can then be tested in computer simulations, where differential equations are solved through massive computation, and the CAD-model is used as input of geometrical data. Further, the CAD data can be used in automatized production. The new technique is revolutionizing the whole industrial process from design to production.

1.5 Navigation: from stars to GPS

A primary force behind the development of geometry and mathematics since the Babylonians has been the need to navigate using information from the positions of the planets, stars, the Moon and the Sun. With a clock and a sextant and mathematical tables, the sea-farer of the 18th century could determine his position more or less accurately. But the results depended strongly on the precision of clocks and observations and it was easy for large errors to creep in. Historically, navigation has not been an easy job.

During the last decade, the classical methods of navigation have been replaced by GPS, the Global Positioning System. With a GPS navigator in hand, which we can buy for a couple of hundred dollars, we get our coordinates (latitude and longitude) with a precision of 50 meters at the press of a button. GPS is based on a simple mathematical principle known already to the Greeks: if we know our distance to three point is space with known coordinates then we can compute our position. The GPS uses this principle by measuring its distance to three satellites with known positions, and then computes its own coordinates. To use this technique, we need to deploy satellites, keep track of them in space and time, and measure relevant distances, which became possible only in the last decades. Of course, computers are used to keep track of the satellites, and the microprocessor of a hand-held GPS measures distances and computes the current coordinates.

The GPS has opened the door to mass consumption in navigation, which was before the privilege of only a few.

1.6 Medical tomography

The computer tomograph creates a pictures of the inside of a human body by solving a certain integral equation by massive computation, with data coming from measuring the attenuation of very weak X-rays sent through the body from different directions. This technique offers mass consumption of medical imaging, which is radically changing medical research and practice.

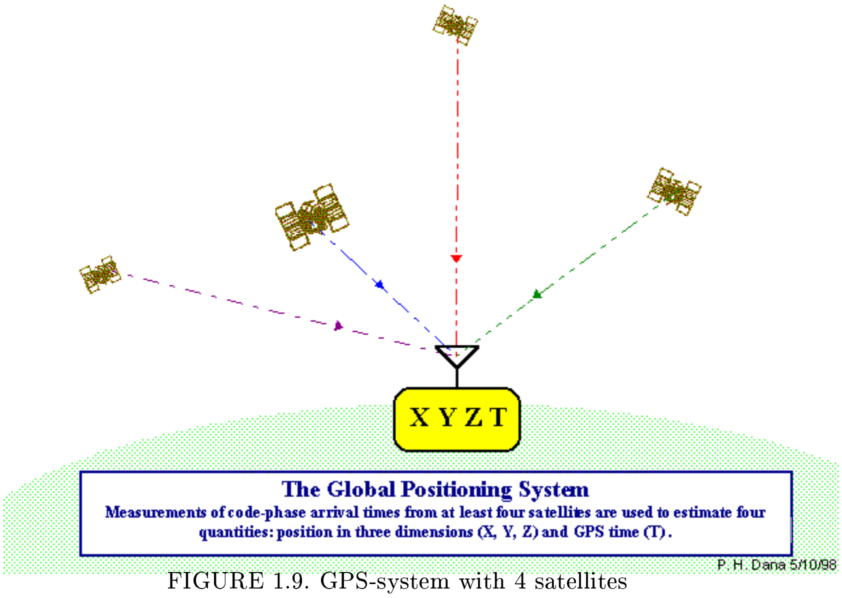


FIGURE 1.9. GPS-system with 4 satellites

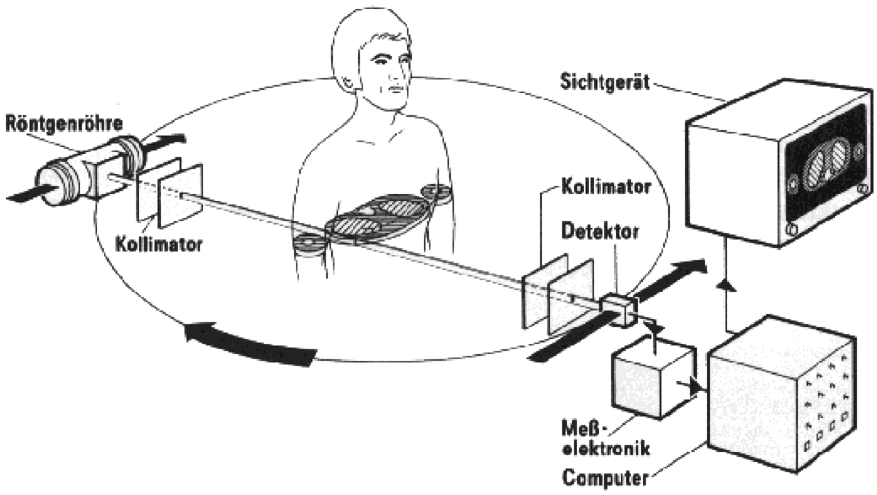


FIGURE 1.10. Medical tomograph.

1.7 Molecular dynamics and medical drug design

The classic way in which new drugs are discovered is an expensive and time-consuming process. First, a physical search is conducted for new organic chemical compounds, for example among the rain forests in South America. Once a new organic molecule is discovered, drug and chemical companies license the molecule for use in a broad laboratory investigation to see if the compound is useful. This search is conducted by expert organic chemists who build up a vast experience with how compounds can interact and which kind of interactions are likely to prove useful for the purpose of controlling a disease or fixing a physical condition. Such experience is needed to reduce the number of laboratory trials that are conducted, otherwise the vast range of possibilities is overwhelming.

The use of computers in the search for new drugs is rapidly increasing. One use is to makeup new compounds so as to reduce the need to make expensive searches in exotic locations like southern rain forests. As part of this search, the computer can also help classify possible configurations of molecules and provide likely ranges of interactions, thus greatly reducing the amount of laboratory testing time that is needed.

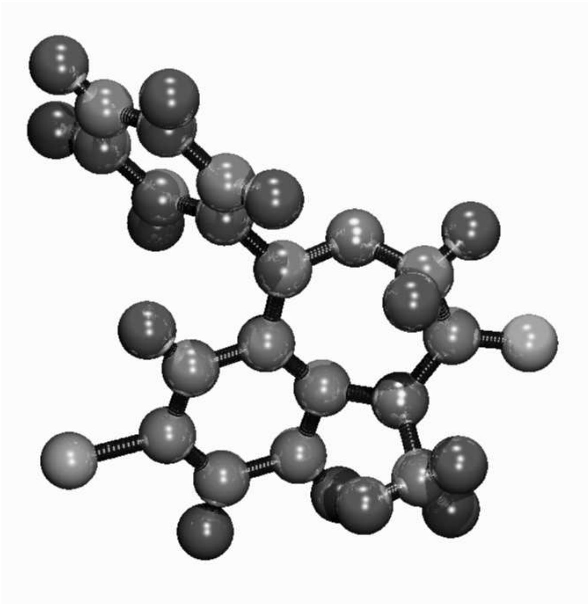


FIGURE 1.11. The Valium molecule

1.8 Weather prediction and global warming

Weather predictions are based on solving differential equations that describe the evolution of the atmosphere using a super computer. Reasonably reliable predictions of daily weather are routinely done for periods of a few days. For longer periods. the reliability of the simulation decreases rapidly, and with present day computers daily weather predictions for a period of two weeks are impossible.

However, forecasts over months of averages of temperature and rainfall are possible with present day computer power and are routinely performed.

Long-time simulations over periods of 20-50 years of yearly temperature-averages are done today to predict a possible *global warming* due to the use of fossil energy. The reliability of these simulations are debated.

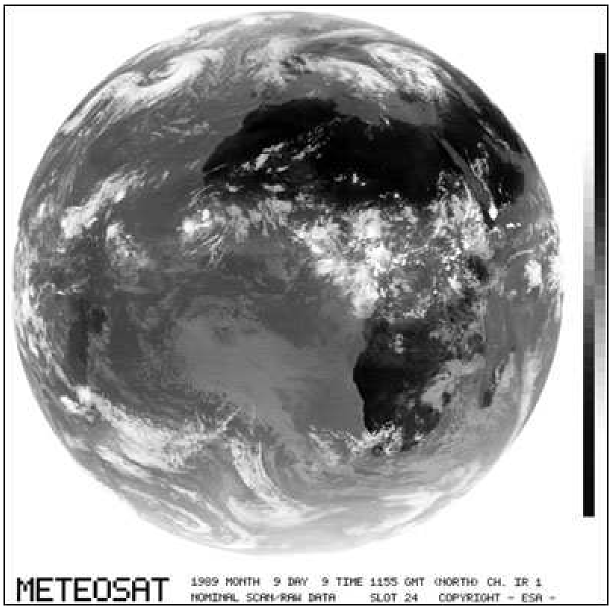


FIGURE 1.12. Clouds over West of Africa (ESA/Eumesat)

1.9 Economy: stocks and options

The Black-Scholes model for pricing options has created a new market of so called derivative trading as a complement to the stock market. To correctly price options is a mathematically complicated and computationally intensive task, and a stock broker with first class software for this purpose (which responds in a few seconds), has a clear trading advantage.

1.10 Languages

Mathematics is a *language*. There are many different languages. Our mother tongue, whatever it happens to be, English, Swedish, Greek, et cetera, is our most important language, which a child masters quite well at the age of three. To learn to write in our native language takes longer time and more effort and occupies a large part of the early school years. To learn to speak and write a foreign language is an important part of secondary education.

Language is used for *communication* with other people for purposes of cooperation, exchange of ideas or control. Communication is becoming increasingly important in our society as the modern means of communication develop.

Using a language we may create *models* of phenomena of interest, and by using models, phenomena may be studied for purposes of *understanding* or *prediction*. Models may be used for *analysis* focussed on a close examination of individual parts of the model and for *synthesis* aimed at understanding the interplay of the parts that is understanding the model as a whole. A *novel* is like a model of the real world expressed in a written language like English. In a novel the characters of people in the novel may be analyzed and the interaction between people may be displayed and studied.

The ants in a group of ants or bees in a bees hive also have a language for communication. In fact in modern biology, the interaction between cells or proteins in a cell is often described in terms of entities "talking to each other".

It appears that we as human beings use our language when we *think*. We then seem to use the language as a model in our head, where we try various possibilities in *simulations* of the real world: "If that happens, then I'll do this, and if instead that happens, then I will do so and so...". Planning our day and setting up our calender is also some type of modeling or simulation of events to come. Simulations by using our language thus seems to go on in our heads all the time.

There are also other languages like the language of musical notation with its notes, bars, scores, et cetera. A musical score is like a model of the real music. For a trained composer, the model of the written score can be very close to the real music. For amateurs, the musical score may say very little, because the score is like a foreign language which is not understood.

1.11 Mathematics as the language of science

Mathematics has been described as the language of science and technology including mechanics, astronomy, physics, chemistry, and topics like fluid and solid mechanics, electromagnetics et cetera. The language of mathematics is used to deal with *geometrical* concepts like *position* and *form* and *mechanical* concepts like *velocity*, *force* and *field*. More generally, mathematics serves as a language in any area that includes *quantitative* aspects described in terms of *numbers*, such as economy, accounting, statistics et cetera. Mathematics serves as the basis for the modern means of electronic *communication* where information is coded as sequences of 0's and 1's and is transferred, manipulated or stored.

The words of the language of mathematics often are taken from our usual language, like *points*, *lines*, *circles*, *velocity*, *functions*, *relations*, *transformations*, *sequences*, *equality*, *inequality* et cetera.

A mathematical word, term or concept is supposed to have a specific meaning defined using other words and concepts that are already defined. This is the same principle as is used in a Thesaurus, where relatively complicated words are described in terms of simpler words. To start the definition process, certain fundamental concepts or words are used, which cannot be defined in terms of already defined concepts. Basic relations between the fundamental concepts may be described in certain *axioms*. Fundamental concepts of Euclidean geometry are *point* and *line*, and a basic Euclidean axiom states that through each pair of distinct points there is a unique line passing. A *theorem* is a statement derived from the axioms or other theorems by using logical reasoning following certain rules of logic. The derivation is called a *proof* of the theorem.

1.12 The basic areas of mathematics

The basic areas of mathematics are

- Geometry
- Algebra
- Analysis.

Geometry concerns objects like *lines*, *triangles*, *circles*. Algebra and Analysis is based on *numbers* and *functions*. The basic areas of mathematics education in engineering or science education are

- Calculus
- Linear Algebra

Calculus is a branch of analysis and concerns properties of functions such as *continuity*, and operations on functions such as *differentiation* and *integration*. Calculus connects to Linear Algebra in the study of *linear functions* or linear transformations and to *analytical geometry*, which describes geometry in terms of numbers. The basic concepts of Calculus are

- function
- derivative
- integral.

Linear Algebra combines Geometry and Algebra and connects to Analytical Geometry. The basic concepts of Linear Algebra are

- vector
- vector space
- projection, orthogonality
- linear transformation

This book teaches the basics of Calculus and Linear Algebra, which are the areas of mathematics underlying most applications.

1.13 What is science?

The theoretical kernel of *natural science* may be viewed as having two components

- formulating equations (modeling),
- solving equations (computation).

Together, these form the essence of *mathematical modeling* and *computational mathematical modeling*. The first really great triumph of science and mathematical modeling is Newton's model of our planetary system as a set of differential equations expressing Newton's law connecting force, through the inverse square law, and acceleration. An *algorithm* may be seen as a strategy or constructive method to solve a given equation via computation. By applying the algorithm and computing, it is possible to simulate real phenomena and make predictions.

Traditional techniques of computing were based on symbolic or numerical computation with pen and paper, tables, slide ruler and mechanical calculator. Automatized computation with computers is now opening new possibilities of simulation of real phenomena according to Natures own principle of massive repetition of simple operations, and the areas of applications are quickly growing in science, technology, medicine and economics.

Mathematics is basic for both steps (i) formulating and (ii) solving equation. Mathematics is used as a language to formulate equations and as a set of tools to solve equations.

Fame in science can be reached by formulating or solving equations. The success is usually manifested by connecting the name of the inventor to the equation or solution method. Examples are legio: Newton's method, Euler's equations, Lagrange's equations, Poisson's equations, Laplace's equations, Navier's equations, Navier-Stokes' equations, Boussinesq's equations, Einstein's equations, Schrödinger's equations, Black-Scholes formula..., some of which we will meet below.

1.14 What is conscience?

The activity of the brain is believed to consist of electrical/chemical signals/waves connecting billions of synapses in some kind of large scale computation. The question of the nature of the *conscience* of human beings has played a central role in the development of human culture since the early Greek civilization, and today computer scientists seek to capture its evasive nature in various forms of Artificial Intelligence AI. The idea of a division of the activity of the brain into a (small) *conscious* "rational" part and a (large) *unconscious* "irrational" part, is widely accepted since the days of Freud. The rational part has the role of "analysis" and "control" towards some "purpose" and thus has features of Soul, while the bulk of the "computation" is Body in the sense that it is "just" electrical/chemical waves. We meet the same aspects in numerical optimization, with the optimization algorithm itself playing the role of Soul directing the computational effort towards the goal, and the underlying computation is Body.

We have been brought up with the idea that the conscious is in control of the mental "computation", but we know that this is often not the case. In fact, we seem to have developed strong skills in various kinds of after-rationalization: whatever happens, unless it is an "accident" or something "unexpected", we see it as resulting from a rational plan of ours made up in advance, thus turning a posteriori observations into a priori predictions.

1.15 How to come to grips with the difficulties of understanding the material of this book and eventually view it as a good friend

We conclude this introductory chapter with some suggestions intended to help the reader through the most demanding first reading of the book and reach a state of mind viewing the book as a good helpful friend, rather than the opposite. From our experience of teaching the material of this book, we know that it may evoke quite a bit of frustration and negative feelings, which is not very productive.

Mathematics is difficult: Choose your own level of ambition.

First, we have to admit that mathematics is a difficult subject, and we see no way around this fact.Secondly, one should realize that it is perfectly possible to live a happy life with a career in both academics and industry with only elementary knowledge of mathematics. There are many examples including Nobel Prize Winners. This means that it is adviceable to set a level of ambition in mathematics studies which is realistic and fits the interest profile of the individual student. Many students of engineering have other prime interests than mathematics, but there are also students who really like mathematics and theoretical engineering subjects using mathematics. The span of mathematical interest thus may be expected to be quite wide in a group of students following a course based on this book, and it seems reasonable that this would be reflected in the choice of level of ambition.

Advanced material: Keep an open mind and be confident.

The book contains quite a bit of material which is "advanced" and not usually met in undergraduate mathematics, and which one may bypass and still be completely happy. It is probably better to be really familiar with and understand a smaller set of mathematical tools and have the ability to meet new challenges with some self-confidence, than repeatedly failing to digest too large portions. Mathematics is so rich, that even a life of fully-time study can only cover a very small part. The most important ability must be to meet new material with an open mind and some confidence!

Some parts of matematics are easy.

On the other hand, there are many aspects of mathematics which are not so difficult, or even "simple", once they have been properly understood. Thus, the book contains both difficult and simple material, and the first impression from the student may give overwhelming weight to the former. To help out we have collected the most essential nontrivial facts in short summaries in the form of *Calculus Tool Bag I and II*, *Linear Algebra Tool Bag*, *Differential Equations Tool Bag*, *Applications Tool Bag*, *Fourier Analysis Tool Bag* and *Analytic Functions Tool Bag*. The reader will find the tool bags surprisingly short: just a couple pages, altogether say 15-20 pages. If properly understood, this material carries a long way and is "all" one needs to remember from the math studies for further studies and professional activities in other areas. Since the book contains about 1200 pages it means

50-100 pages of book text for each one page of summary. This means that the book gives more than the absolute minimum of information and has the ambition to give the mathematical concepts a perspective concerning both history and applicability today. So we hope the student does not get turned off by the quite a massive number of words, by remembering that after all 15-20 pages captures the essential facts. During a period of study of say one year and a half of math studies, this effectively means about one third of a page each week!

Increased/decreased importance of mathematics

The book reflects both the increased importance of mathematics in the information society of today, and the decreased importance of much of the analytical mathematics filling the traditional curriculum. The student thus should be happy to know that many of the traditional formulas are no longer such a must, and that a proper understanding of relatively few basic mathematical facts can help a lot in coping with modern life and science.

Which chapters can I skip in a first reading?

We indicate by * certain chapters directed to applications, which one may by-pass in a first reading without loosing the main thread of the presentation, and return to at a later stage if desired.

Chapter 1 Problems

1.1. Find out which Nobel Prize Winners got the prize for formulating or solving equations.

1.2. Reflect about the nature of “thinking” and “computing”.

1.3. Find out more about the topics mentioned in the text.

1.4. (a) Do you like mathematics or hate mathematics, or something in between? Explain your standpoint. (b) Specify what you would like to get out of your studies of mathematics.

1.5. Present some basic aspects of science.

FIGURE 1.13. Left person: "Isn't it remarkable that one can compute the distance to stars like Cassiopeja, Aldebaran and Sirius?". Right person: "I find it even more remarkable that one may know their names!" (Assar by Ulf Lundquist)

2
The Mathematics Laboratory

2.1 Introduction

This book is complemented by various pieces of software collected into the *Mathematics Laboratory*, freely available on the book web site. The *Mathematics Laboratory* contains different types of software organized under the following headings:

- *Math Experience*
- *Tools*
- *Applications*
- *Students Lab*

Math Experience serves the purpose of illustrating mathematical concepts from analysis and linear algebra such as roots of equations, Lipschitz continuity, fixed point iteration, differentiability, the definition of the integral, basis calculus for functions of several varibles.

Tools contains (i) *ready-mades* such as different solvers for differential equations and (ii) *shells* aimed at helping the student to make his own tools such as solvers of systems of equations and differential equations.

Applications contains more developed sowtware like *DOLFIN*, a platform for Dynamic Objectoriented Library for FINite Elements, *Tanganyika* for multiadaptive solution of systems of differential equations and *Mech-lab* and for simulating mechanical systems.

Many students may feel stimulated to develop their own applications during a mathematics program based on this book using the basis offered by the *Mathematics Laboratory*. We invite students to contribute applications to the *Students Lab*, which we hope may serve as a source of inspiration transferring know how from old to new students.

2.2 Math Experience

Math Experience is a collection of Matlab GUI software designed to offer a deeper understanding of important mathematical concepts and ideas such as, for example, convergence, continuity, linearization, differentiation, Taylor polynomials, integration, etc. The idea is to provide on-screen computer "labs" in which the student, by himself guided by a number of well designed questions, can seek to fully understand (a) the concepts and ideas as such and (b) the mathematical formulas and equations describing the concepts, by interacting with the lab environment in different ways. For example, in the Taylor lab (see Fig. 2.1) it is possible to give a function, or pick one from a gallery, and study its Taylor polynomial approximation of different degrees, how it depends on the point of focus by mouse-dragging the point, how it depends on the distance to the point by zooming in and out etc. There is also a movie where the terms in the Taylor polynomial are added one at a time. In the MultiD Calculus lab (see Fig. 2.2) it is possible to define a function $u(x_1, x_2)$ and compute its integral over a given curve or a given domain, to view its gradient field, contour plots, tangent planes etc. One may also study vector fields (u, v), view their divergence and rotation, compute the integrals of these quantities to verify the fundamental theorems of vector calculus, view the (u, v) mapped domain and the Jacobian of the map etc, etc.

Several labs of this type have been completed and will be made available from the book web page (see www.phi.chalmers.se), and more are under way according to the following list:

- Func lab - about relations and functions, inverse function etc.
- Graph Gallery - elementar functions and their parameter dependence.
- Cauchy lab - about sequences & convergence
- Lipschitz lab - the concept of continuity
- Root lab - about bisection and fixed point iteration
- Linearization and the derivative
- Newtons lab - illustrating Newtons method

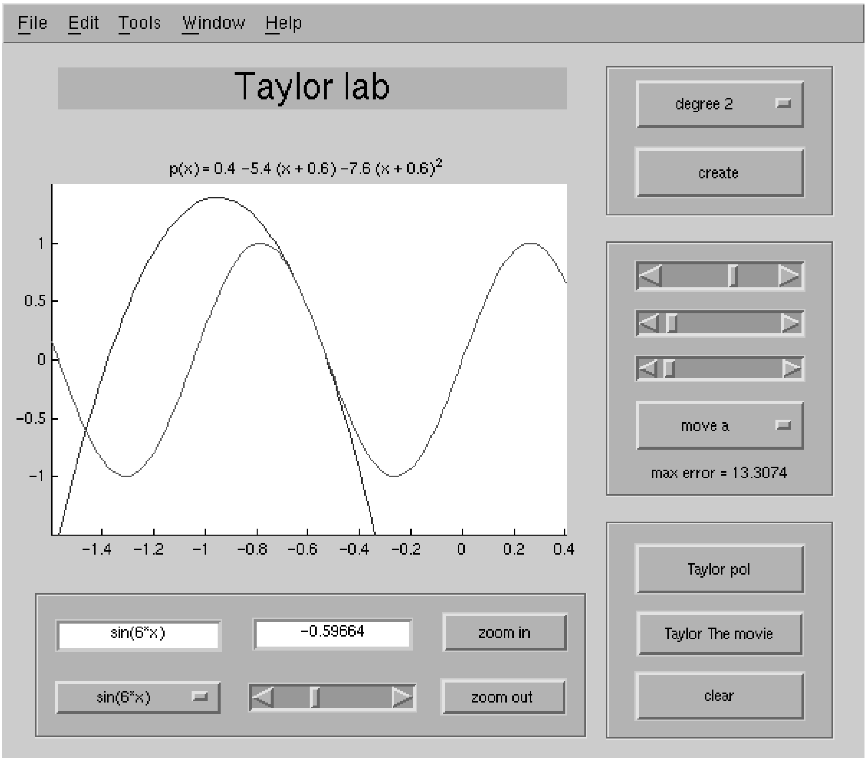


FIGURE 2.1. The Taylor lab

- Taylor lab - polynomials
- Opti lab - elementary optimization
- Piecewise polynomial lab - about piecewise polynomial approximation
- Integration lab - Euler and Riemann summation, adaptive integration
- Dynamical system lab
- Pendulum lab - the effect of linearization and approximation
- Vector algebra - a graphical vector calculator
- Analytic geometry - coupling geometry and linear algebra
- MultiD Calculus - integration and vector Calculus
- FE-lab - illustrating the finite element method
- Adaptive FEM - illustrating adaptive finite element techniques

- Poisson lab - fundamental solutions etc
- Fourier lab
- Wavelet lab
- Optimal control lab - control problems related to differential equations
- Archimedes lab - for experiments related to Archimedes principle

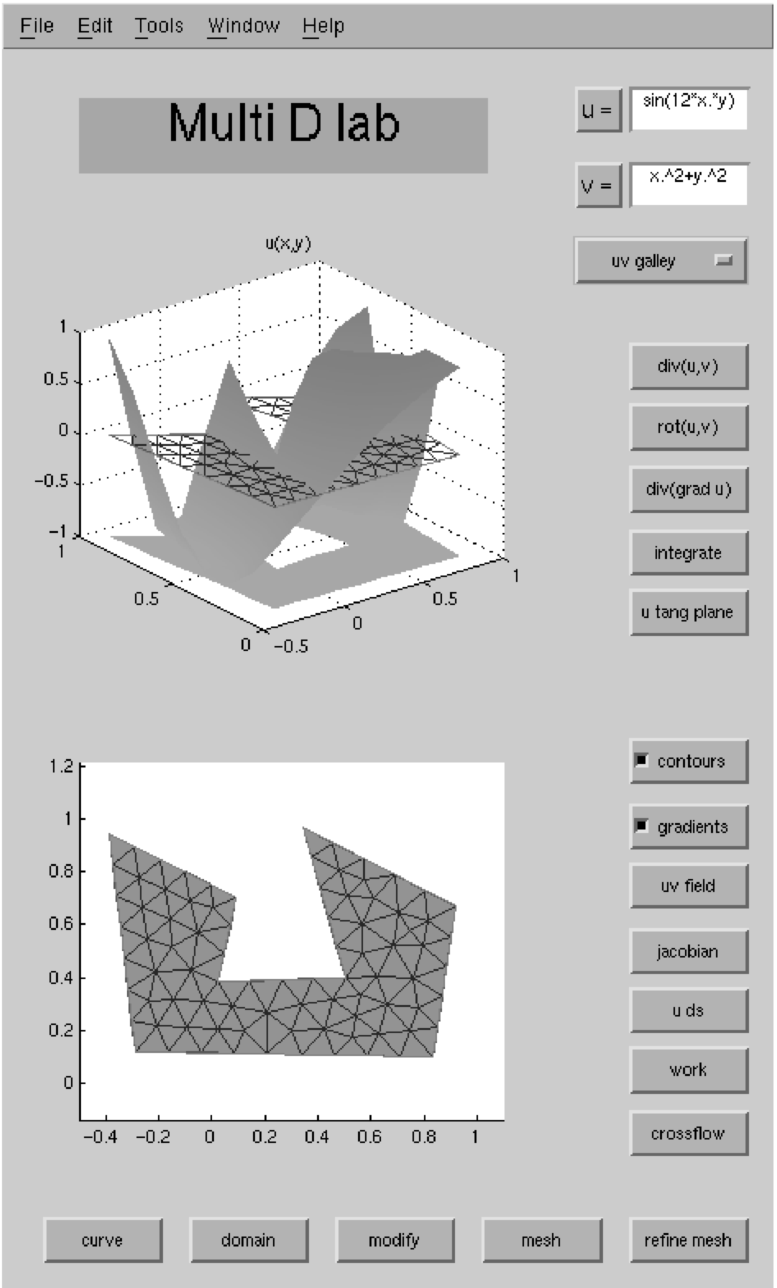


FIGURE 2.2. The MultiD Calculus lab

3
Introduction to modeling

> The best material model of a cat is another, or preferably the same, cat. (Rosenblueth/Wiener in Philosophy of Science 1945)

3.1 Introduction

We start by giving two basic examples of the use of mathematics for describing practical situations. The first example is a problem in household economy and the second is a problem in surveying, both of which have been important fields of application for mathematics since the time of the Babylonians. The models are very simple but illustrate fundamental ideas.

3.2 The Dinner Soup model

You want to make a soup for dinner together with your roommate, and following a recipe you ask your roommate to go to the grocery store and buy 10 dollars worth of potatoes, carrots, and beef according to the proportions 3:2:1 by weight. In other words, your roommate has 10 dollars to spend on the ingredients, which should be bought in the amounts so that by weight there are three times as much potatoes as beef and two times as much carrots as beef. At the grocery store, your roommate finds that potatoes are 1 dollar per pound, carrots are 2 dollars per pound, and beef is 8 dollars per pound. Your roommate thus faces the problem of figuring out how much of each ingredient to buy to use up the 10 dollars.

One way to solve the problem is by trial and error as follows: Your roommate could take quantities of the ingredients to the cash register in the proportions of 3:2:1 and let the clerk check the price, repeating until a total of 10 dollars is reached. Of course, both your roommate and the clerk could probably think of better ways to spend the afternoon. Another possibility would be to make a *mathematical model* of the situation and then seek to find the correct amounts to buy by doing some computations. The basic idea would be to use brains and pen and paper or a calculator, instead of labor intensive brute physical work.

The mathematical model may be set up as follows: Recalling that we want to determine the amounts of ingredients to buy, we notice that it is enough to determine the amount of beef, since we'll buy twice as much carrots as beef and three times as much potatoes as beef. Let's give a name to the quantity to determine. Let x denote the amount of meat in pounds to buy. The *symbol* x here represents an unknown quantity, or *unknown*, that we are seeking to determine by using available information.

If the amount of meat is x pounds, then the price of the meat to buy is $8x$ dollars by the simple computation

$$\text{cost of meat in dollars} = x\,\text{pounds} \times 8\,\frac{\text{dollars}}{\text{pound}} = 8x\,\text{dollars}.$$

Since there should be three times as much potatoes as meat by weight, the amount of potatoes in pounds is $3x$ and the cost of the potatoes is $3x$ dollars since the price of potatoes is one dollar per pound. Finally, the amount of carrots to buy is $2x$ and the cost is 2 times $2x = 4x$ dollars, since the price is 2 dollars per pound. The total cost of meat, potatoes and carrots is found by summing up the cost of each

$$8x + 3x + 4x = 15x.$$

Since we assume that we have 10 dollars to spend, we get the relation

$$15x = 10, \tag{3.1}$$

which expresses the equality of total cost and available money. This is an *equation* involving the unknown x and data determined by the physical situation. From this equation, your roommate can figure out how much beef to buy. This is done by dividing both sides of (3.1) by 15, which gives $x = 10/15 = 2/3 \approx 0.67$ pounds of meat. The amount of carrots should then be $2\times 2/3 = 4/3 \approx 1.33$, and finally the amount of potatoes $3\times 2/3 = 2$ pounds.

The *mathematical model* for this situation is $15x = 10$, where x is the amount of meat, $15x$ is the total cost and 10 is the available money. The modeling consists in expressing the total cost of the ingredients $15x$ in terms of the amount of beef x. Note that in this model, we only take into account what is essential for the current purpose of buying potatoes, carrots and meat for the Dinner Soup, and we did not bother to write down the prices of other items, like ice cream or beer. Determining the useful information is an important, and sometimes difficult, part of the mathematical modeling.

A nice feature of mathematical models is that they can be reused to simulate different situations. For example, if you have 15 dollars to spend, then the model $15x = 15$ arises with solution $x = 1$. If you have 25 dollars to spend, then the model is $15x = 25$ with solution $x = 25/15 = 5/3$. In general, if the amount of money y is given, then the model is $15x = y$. In this model we use the two symbols x and y, and assume that the amount of money y is given and the amount of beef x is an unknown quantity to be determined from the equation ($15x = y$) of the model. The roles could shift around: you may think of the amount of beef x as being given and the total cost or expenditure y to be determined (according to the formula $y = 15x$). In the first case, we would think of the amount of beef x as a function of the expenditure y and in the second the expenditure y as a function of x.

Assigning symbols to relevant quantities, known or unknown, is an important step in setting up a mathematical model of something. The idea of assigning symbols for unknown quantities was used already by the Babylonians (who had frequent use of models like the Dinner Soup model in organizing the feeding of the many people working on their irrigation systems).

Suppose that we could not solve the equation $15x = 10$, because of a lack of skill in solving equations (we may have forgotten the trick of dividing by 15 that we learned in school). We could then try to get a solution by some kind of trial and error strategy as follows. First we assume that $x = 1$. We then find that the total cost is 15 dollars, which is too much. We then try with a smaller quantity of meat, say $x = 0.6$, and compute the total cost to 9 dollars, which is too little. We then try with something between 0.6 and 1, say $x = 0.7$ and find that the cost would be 10.5 dollars, which is a little too much. We conclude that the right amount must be somewhere between 0.6 and 0.7, probably closer 0.7. We can continue in the same way to find as many decimals of x as we like. For instance we check next in the same way that x must be some where between 0.66 and 0.67. In this case we know the exact answer $x = \frac{2}{3} = 0.66666....$ The trial and error strategy just described is a model of the process of bringing food to the counter and letting the cashier compute the total prize. In the model we compute the prize ourselves without having to physically collect the items and bring them to the counter, which simplifies the trial and error process.

3.3 The Muddy Yard model

One of the authors owns a house with a 100m × 100m backyard that has the unfortunate tendency to form a muddy lake every time it rains. We show a perspective of the field on the left in Fig. 3.1. Because of the grading in

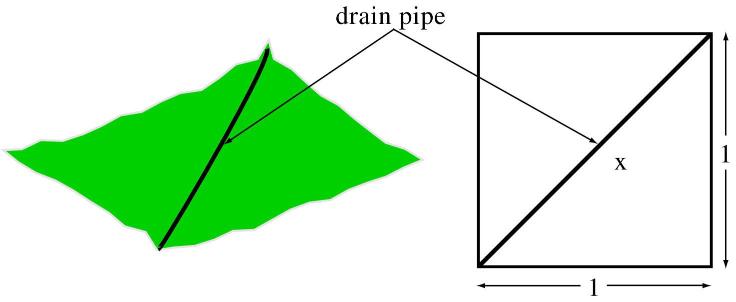


FIGURE 3.1. Perspective of a field with poor drainage and a model describing the dimensions.

the yard, the owner has had the idea for some time to fix this by digging a shallow ditch down the diagonal of the yard, laying some perforated plastic drain pipe, then covering the pipe back up. He is then faced with the problem of determining the amount of pipe that he needs to purchase. Since a survey of the property only provides the outside dimensions and the locations of corners and physically measuring the diagonal through the mud is not easy to do, he has decided to try to compute the distance using mathematics. Can mathematics help him in this endeavor?

Inspection of the property and a map indicates that the yard can be modeled as a horizontal square (the grading seems small), and we thus seek to compute the length of the diagonal of the square. We display the model on the right in Fig. 3.1, where we change to units of 100m, so the field is 1×1, and denote the length of the diagonal by x. We now recall Pythagoras' theorem, which states that $x^2 = 1^2 + 1^2 = 2$. To find the length x of the drain pipe, we are thus led to solve the equation

$$x^2 = 2. \tag{3.2}$$

Solving the equation $x^2 = 2$ may seem to be deceptively simple at first; the positive solution is just $x = \sqrt{2}$ after all. But walking into a store and asking for $\sqrt{2}$ units of pipe may not get a positive response. Precut pipes do not come in lengths calibrated by $\sqrt{2}$, neither do measure sticks indicate $\sqrt{2}$, and a clerk is thus going to need some concrete information about the value of $\sqrt{2}$ to be able to measure out a proper piece of pipe.

We can try to pin down the value of $\sqrt{2}$ by using a trial and error strategy. We can check easily that $1^2 = 1 < 2$ while $2^2 = 4 > 2$. So we know that

$\sqrt{2}$, whatever it is, is between 1 and 2. Next we can check $1.1^2 = 1.21$, $1.2^2 = 1.44$, $1.3^2 = 1.69$, $1.4^2 = 1.96$, $1.5^2 = 2.25$, $1.6^2 = 2.56$, $1.7^2 = 2.89$, $1.8^2 = 3.24$, $1.9^2 = 3.61$. Apparently $\sqrt{2}$ is between 1.4 and 1.5. Next we can try to fix the third decimal. Now we find that $1.41^2 = 1.9881$ while $1.42^2 = 2.0164$. So apparently $\sqrt{2}$ is between 1.41 and 1.42 and likely closer to 1.41. It appears that proceeding in this way, we can determine as many decimal places of $\sqrt{2}$ as we like, and we may consider the problem of computing how much drain pipe to buy to be solved!

Below we will meet many equations that have to be solved by using some variation of a trial and error strategy. In fact, most mathematical equations cannot be solved exactly by some algebraic manipulations, as we could do (if we were sufficiently clever) in the case of the Dinner Soup model (3.1). Consequently, the trial and error approach to solving mathematical equations is fundamentally important in mathematics. We shall also see that trying to solve equations such as $x^2 = 2$ carries us directly into the very heart of mathematics, from Pythagoras and Euclid through the quarrels on the foundations of mathematics that peaked in the 1930s and on into the present day of the modern computer.

3.4 A system of equations: Dinner Soup/Ice Cream model

Suppose you would like to finish off the Dinner Soup with some ice cream dessert at the cost of 3 dollars a pound, still at the total expense of 10 dollars. How much of each item should now be bought?

Well, if the amount ice cream is y pounds, the total cost will be $15x + 3y$ and thus we have the equation $15x + 3y = 10$ expressing that the total cost is equal to the available money. We now have two unknowns x and y, and we need one more equation. So far, we would be able to set $x = 0$ and solve for $y = \frac{10}{3}$ spending all the money on ice cream. This would go against some principle we learned as small kids. The second equation needed could come from some idea of balancing the amount of ice cream (junk food) to the amount of carrots (healthy food), for example according to the formula $2x = y + 1$, or $2x - y = 1$. Altogether, we would thus get the following system of two equations in the two unknowns x and y:

$$\begin{aligned} 15x + 3y &= 10, \\ 2x - y &= 1. \end{aligned}$$

Solving for y in the second equation, we get $y = 2x - 1$, which inserted into the first equation gives

$$15x + 6x - 3 = 10, \quad \text{that is} \quad 21x = 13, \quad \text{that is,} \quad x = \frac{13}{21}.$$

Finally, inserting the value of $x = \frac{13}{21} \approx 0.60$ into the equation $2x - y = 1$, we get $y = \frac{5}{21} \approx 0.24$, and we have found the solution of the system of equations modeling the present situation.

3.5 Formulating and solving equations

Let us put the Dinner Soup and Muddy Yard models into the perspective of formulating and solving equations presented above. Formulating equations corresponds to collecting information in systematic form, and solving equations corresponds to drawing conclusions from the collected information.

We began by describing the physical situations in the Dinner Soup and Muddy Yard models in terms of mathematical equations. This aspect is not just mathematical but involves also whatever knowledge from physics, economy, history, psychology, etcetera that may be relevant to describe the situation to be modeled. The equations we obtained in the Dinner Soup and the Muddy Yard models, namely $15x = 10$ and $x^2 = 2$, are examples of *algebraic equations* in which the data and the unknown x are both numbers. As we consider more complicated situations, we will often encounter models in which the data and the unknown quantities are *functions*. Such models typically contain derivatives and integrals and are then referred to as *differential equations* or *integral equations*.

The second aspect is *solving* the equations of the model to determine the unknown and gain new information about the situation at hand. In the case of the Dinner Soup model, we can solve the model equation $15x = 10$ exactly and express the solution $x = 2/3$ as a rational number. In the Muddy Yard model we resort to an iterative "trial and error" strategy to compute as many digits of the decimal expansion of the solution $x = \sqrt{2}$ as we may need. So it goes in general: once in a while, we can write down a solution of a model equation explicitly, but most often we have to be content with an approximate solution, the digits of which we can determine through some iterative computational process.

Note that there is no reason to be disappointed over the fact that equations representing mathematical models cannot be solved exactly, since the mathematical model is an approximation anyway, in general. It is better to have a complicated but accurate mathematical model equation, that admits only approximate solutions, than to have a trivial inaccurate model equation that can be solved exactly! The wonderful thing with computers is that they may compute accurate solutions also to complicated accurate model equations, which makes it possible to simulate real phenomena.

Chapter 3 Problems

3.1. Suppose that the grocery store sells potatoes for 40 cents per pound, carrots for 80 cents per pound, and beef for 40 cents per *ounce*. Determine the model relation for the total price.

3.2. Suppose that you change the soup recipe to have equal amounts of carrots and potatoes while the weight of these combined should be six times the weight of beef. Determine the model relation for the total price.

3.3. Suppose you go all out and add onions to the soup recipe in the proportion of 2 : 1 to the amount of beef, while keeping the proportions of the other ingredients the same. The price of onions in the store is \$1 per pound. Determine the model relation for the total price.

3.4. While flying directly over the airport in a holding pattern at an altitude of 1 mile, you see your high rise condominium from the window. Knowing that the airport is 4 miles from your condominium and pretending that the condominium has height 0, how far are you from home and a cold beer?

3.5. Devise a model of the draining of a yard that has three sides of approximately the same length 2 assuming that we drain the yard by laying a pipe from one corner to the midpoint of the opposite side. What quantity of pipe do we need?

3.6. A father and his child are playing with a teeter-totter which has a seatboard 12 feet long. If the father weighs 170 pounds and the child weighs 45 pounds, construct a model for the location of the pivot point on the board in order for the teeter-totter to be in perfect balance? Hint: recall the principle of the lever which says that the products of the distances from the fulcrum to the masses on each end of a lever must be equal for the lever to be in equilibrium.

4

A very short Calculus course

> Mathematics has the completely false reputation of yielding infallible conclusions. Its infallibility is nothing but identity. Two times two is not four, but it is just two times two, and that is what we call four for short. But four is nothing new at all. And thus it goes on in its conclusions, except that in the height the identity fades out of sight. (Goethe)

4.1 Introduction

Following up on the general idea of science as a combination of formulating and solving equations, we describe the bare elements of this picture from a mathematical point of view. We want to give a brief glimpse of the main themes of Calculus that will be discovered as we work through these books. In particular, we will encounter the magical words of *function*, *derivative*, and *integral*. If you have some idea of these concepts already, you will understand some of the outline. If you have no prior acquaintance with these concepts, you can use this section to just get a first taste of what Calculus is all about without expecting to understand the details at this point. Keep in mind that this is just a glimpse of the actors behind the curtain before the play begins!

We hope the reader can use this chapter to get a grip on the essence of Calculus by reading just a couple of pages. But this is really impossible in some sense because calculus contains so many formulas and details that it is easy to get overwhelmed and discouraged. Thus, we urge the reader to browse through the following couple of pages to get a quick idea and then return later and confirm with an “of course”.

On the other hand, the reader may be surprised that something that is seemingly explained so easily in a couple of pages, actually takes several hundred pages to unwind in this book (and other books). We don’t seem to be able give a good explanation of this “contradiction” indicating that “what looks difficult may be easy” and vice versa. We also present short summaries of Calculus in Chapter *Calculus Tool Bag I Calculus Tool Bag II*, which support the idea that a destillated essence of Calculus indeed can be given in a couple of pages.

4.2 Algebraic equations

We will consider *algebraic equations* of the form: find $\bar{x}$ such that

$$f(\bar{x}) = 0, \tag{4.1}$$

where $f(x)$ is a *function* of x. Recall that $f(x)$ is said to be a function of x if for each number x there is a number $y = f(x)$ assigned. Often, $f(x)$ is given by some algebraic formula: for example $f(x) = 15x - 10$ as in the Dinner Soup model, or $f(x) = x^2 - 2$ as in the Muddy Yard model.

We call $\bar{x}$ a *root* of the equation $f(x) = 0$ if $f(\bar{x}) = 0$. The root of the equation $15x - 10 = 0$ is $\bar{x} = \frac{2}{3}$. The positive root $\bar{x}$ of the equation $x^2 - 2 = 0$ is equal to $\sqrt{2} \approx 1.41$. We will consider different methods to compute a root $\bar{x}$ satisfying $f(\bar{x}) = 0$, including the trial and error method briefly presented above in the context of the Muddy Yard Model.

We will also meet *systems of algebraic equations*, where we seek to determine several unknowns satisfying several equations, as for the Dinner Soup/Ice cream model above.

4.3 Differential equations

We will also consider the following *differential equation*: find a function $x(t)$ such that for all t

$$x'(t) = f(t), \tag{4.2}$$

where $f(t)$ is a given function, and $x'(t)$ is the *derivative* of the function $x(t)$. This equation has several new ingredients. First, we seek here a *function* $x(t)$ with a set of different values $x(t)$ for different values of the variable t, and not just one single value of x like the root the algebraic equation $x^2 = 2$ considered above. In fact, we met this already in the Dinner Soup problem in case of a variable amount of money y to spend, leading to the equation $15x = y$ with solution $x = \frac{y}{15}$ depending on the variable y, that is,

$x = x(y) = \frac{y}{15}$. Secondly, the equation $x'(t) = f(t)$ involves the derivative $x'(t)$ of $x(t)$, so we have to investigate derivatives.

A basic part of Calculus is to (i) explain what a derivative is, and (ii) solve the differential equation $x'(t) = f(t)$, where $f(t)$ is a given function. The solution $x(t)$ of the differential equation $x'(t) = f(t)$, is referred to as an *integral* of $f(t)$, or alternatively as a *primitive function* of $f(t)$. Thus, a basic problem of Calculus is to find a primitive function $x(t)$ of a given function $f(t)$ corresponding to solving the differential equation $x'(t) = f(t)$.

We now attempt to explain (i) the meaning of (4.2) including the meaning of the derivative $x'(t)$ of the function $x(t)$, and (ii) give a hint at how to find the solution $x(t)$ of the differential equation $x'(t) = f(t)$ in terms of the given function $f(t)$.

As a concrete illustration, let us imagine a car moving on a highway. Let t represent *time*, let $x(t)$ be the *distance* traveled by the car at time t, and let $f(t)$ be the *momentary velocity* of the car at time t, see Fig. 4.1.

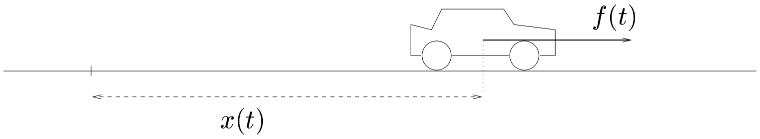


FIGURE 4.1. Highway with car (Volvo?) with velocity $f(t)$ and travelled distance $x(t)$.

We choose a starting time, say $t = 0$ and a final time, say $t = 1$, and we watch the car as it passes from its initial position with $x(0) = 0$ at time $t = 0$ through a sequence of increasing intermediate times t_1, t_2,..., with corresponding distances $x(t_1)$, $x(t_2)$,..., to the final time $t = 1$ with total distance $x(1)$. We thus assume that $0 = t_0 < t_1 < \cdots < t_{n-1} < t_n \cdots < t_N = 1$ is a sequence of intermediate times with corresponding distances $x(t_n)$ and velocities $f(t_n)$, see Fig. 4.2

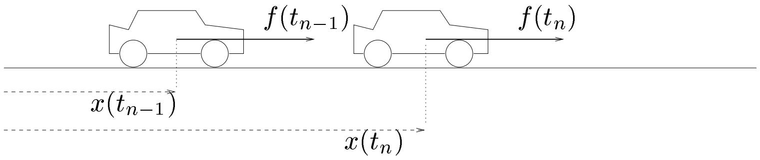


FIGURE 4.2. Distance and velocity at times t_{n-1} and t_n.

For two consecutive times t_{n-1} and t_n, we expect to have

$$x(t_n) \approx x(t_{n-1}) + f(t_{n-1})(t_n - t_{n-1}), \tag{4.3}$$

which says that the distance $x(t_n)$ at time t_n is obtained by adding to the distance $x(t_{n-1})$ at time t_{n-1} the quantity $f(t_{n-1})(t_n - t_{n-1})$, which is the product of the velocity $f(t_{n-1})$ at time t_{n-1} and the *time increment* $t_n - t_{n-1}$. This is because

$$\text{change in distance} \;=\; \text{average velocity} \;\times\; \text{change in time},$$

or traveled distance between time t_{n-1} and t_n equals the (average) velocity multiplied by the time change $t_n - t_{n-1}$. Note that we may equally well connect $x(t_n)$ to $x(t_{n-1})$ by the formula

$$x(t_n) \approx x(t_{n-1}) + f(t_n)(t_n - t_{n-1}), \tag{4.4}$$

corresponding to replacing t_{n-1} by t_n in the $f(t)$-term. We use the approximate equality $\approx$ because we use the velocity $f(t_{n-1})$ or $f(t_n)$, which is not exactly the same as the average velocity over the time interval from t_{n-1} to t_n, but should be close to the average if the time interval is short.

EXAMPLE 4.1. If $x(t) = t^2$, then $x(t_n) - x(t_{n-1}) = t_n^2 - t_{n-1}^2 = (t_n + t_{n-1})(t_n - t_{n-1})$, and (4.3) and (4.4) correspond to approximating the average velocity $(t_n + t_{n-1})$ with $2t_{n-1}$ or $2t_n$, respectively.

The formula (4.3) is at the heart of Calculus! It contains both the derivative of $x(t)$ and the integral of $f(t)$. First, shifting $x(t_{n-1})$ to the left and then dividing by the time increment $t_n - t_{n-1}$, we get

$$\frac{x(t_n) - x(t_{n-1})}{t_n - t_{n-1}} \approx f(t_{n-1}). \tag{4.5}$$

This is a counterpart to (4.2), which indicates how to define the derivative $x'(t_{n-1})$ in order to have the equation $x'(t_{n-1)}) = f(t_{n-1})$ fulfilled:

$$x'(t_{n-1}) \approx \frac{x(t_n) - x(t_{n-1})}{t_n - t_{n-1}}. \tag{4.6}$$

This formula says that the derivative $x'(t_{n-1})$ is approximately equal to the *average velocity*

$$\frac{x(t_n) - x(t_{n-1})}{t_n - t_{n-1}}.$$

over the time interval between t_{n-1} and t_n. Thus, we may expect that the equation $x'(t) = f(t)$ just says that *the derivative $x'(t)$ of the traveled distance $x(t)$ with respect to time t, is equal to the momentary velocity $f(t)$*. The formula (4.6) then says that the velocity $x'(t_{n-1})$ at time t_{n-1}, that is the *momentary velocity* at time t_{n-1}, is approximately equal to the *average velocity* over the time interval (t_{n-1}, t_n). We have now uncovered some of the mystery of the derivative hidden in (4.3).

Next, considering the formula corresponding to to (4.3) for the time instances t_{n-2} and t_{n-1}, obtained by simply replacing n by $n-1$ everywhere in (4.3), we have

$$x(t_{n-1}) \approx x(t_{n-2}) + f(t_{n-2})(t_{n-1} - t_{n-2}), \tag{4.7}$$

and thus together with (4.3),

$$x(t_n) \approx \overbrace{x(t_{n-2}) + f(t_{n-2})(t_{n-1} - t_{n-2})}^{\approx x(t_{n-1})} + f(t_{n-1})(t_n - t_{n-1}). \tag{4.8}$$

Repeating this process, and using that $x(t_0) = x(0) = 0$, we get the formula

$$\begin{aligned} x(t_n) \approx X_n &= f(t_0)(t_1 - t_0) + f(t_1)(t_2 - t_1) + \cdots \\ &\quad + f(t_{n-2})(t_{n-1} - t_{n-2}) + f(t_{n-1})(t_n - t_{n-1}). \end{aligned} \tag{4.9}$$

EXAMPLE 4.2. Consider a velocity $f(t) = \frac{t}{1+t}$ increasing with time t from zero for $t = 0$ towards one for large t. What is the travelled distance $x(t_n)$ at time t_n in this case? To get an (approximate) answer we compute the approximation X_n according to (**??**):

$$\begin{aligned} x(t_n) \approx X_n &= \frac{t_1}{1+t_1}(t_2 - t_1) + \frac{t_2}{1+t_2}(t_3 - t_2) + \cdots \\ &+ \frac{t_{n-2}}{1+t_{n-2}}(t_{n-1} - t_{n-2}) + \frac{t_{n-1}}{1+t_{n-1}}(t_n - t_{n-1}). \end{aligned}$$

With a "uniform" time step $k = t_j - t_{j-1}$ for all j, this reduces to

$$\begin{aligned} x(t_n) \approx X_n &= \frac{k}{1+k}k + \frac{2k}{1+2k}k + \cdots \\ &+ \frac{(n-2)k}{1+(n-2)k}k + \frac{(n-1)k}{1+(n-1)k}k. \end{aligned}$$

We compute the sum for $n = 1, 2, .., N$ choosing $k = 0.05$, and plot the resulting values of X_n inFig. 4.3.

We now return to (4.9), and setting $n = N$ we have in particular

$$\begin{aligned} x(1) = x(t_N) &\approx f(t_0)(t_1 - t_0) + f(t_1)(t_2 - t_1) + \cdots \\ &+ f(t_{N-2})(t_{N-1} - t_{N-1}) + f(t_{N-1})(t_N - t_{N-1}), \end{aligned}$$

that is, $x(1)$ is (approximately) the sum of the terms $f(t_{n-1})(t_n - t_{n-1})$ with n ranging from $n = 1$ up to $n = N$. We may write this in more condensed form using the *summation sign* Σ as

$$x(1) \approx \sum_{n=1}^{N} f(t_{n-1})(t_n - t_{n-1}), \tag{4.10}$$

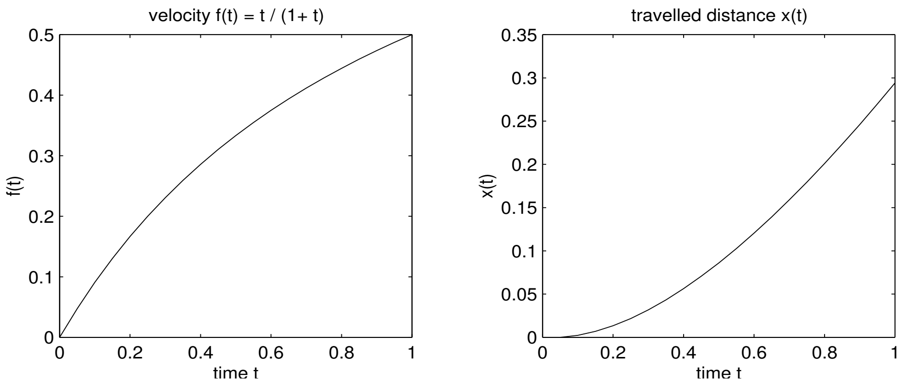


FIGURE 4.3. Travelled distance X_n approximating $x(t)$ for $f(t) = \frac{t}{1+t}$ with time steps $k = 0.05$.

which expresses the total distance $x(1)$ as the sum of all the increments of distance $f(t_{n-1})(t_n - t_{n-1})$ for $n = 1, ..., N$. We can view this formula as a variant of the "telescoping" formula

$$\begin{aligned} x(1) &= x(t_N) \overbrace{-x(t_{N-1}) + x(t_{N-1})}^{=0} \overbrace{-x(t_{N-2}) + x(t_{N-2})}^{=0} \cdots \\ &\qquad + \overbrace{-x(t_1) + x(t_1)}^{=0} - x(t_0) \\ &= \underbrace{x(t_N) - x(t_{N-1})}_{\approx f(t_{N-1})(t_N - t_{N-1})} + \underbrace{x(t_{N-1}) - x(t_{N-2})}_{\approx f(t_{N-2})(t_{N-1} - t_{N-2})} + x(t_{N-2}) \cdots - x(t_1) \\ &\qquad + \underbrace{x(t_1) - x(t_0)}_{\approx f(t_0)(t_1 - t_0)} \end{aligned}$$

expressing the total distance $x(1)$ as a sum of all the increments $x(t_n) - x(t_{n-1})$ of distance (assuming $x(0) = 0$), and recalling that

$$x(t_n) - x(t_{n-1}) \approx f(t_{n-1})(t_n - t_{n-1}).$$

In the telescoping formula, each value $x(t_n)$, except $x(t_N) = x(1)$ and $x(t_0) = 0$, occurs twice with different signs.

In the language of Calculus, the formula (4.10) will be written as

$$x(1) = \int_0^1 f(t)\, dt, \tag{4.11}$$

where $\approx$ has been replaced by $=$, the sum $\sum$ has been replaced by the *integral* $\int$, the increments $t_n - t_{n-1}$ by dt, and the sequence of "discrete" time instances t_n, running (or rather "jumping" in small steps) from time 0 to time 1 corresponds to the *integration variable* t running ("continuously") from 0 to 1. We call the right hand side of (4.11) the *integral* of $f(t)$ from 0 to 1. The value $x(1)$ of the function $x(t)$ for $t = 1$, is the *integral* of $f(t)$ from 0 to 1. We have now uncovered some of the mystery of the integral hidden in the formula (4.10) resulting from summing the basic formula (4.3).

The difficulties with Calculus, in short, are related to the fact that in (4.5) we divide with a small number, namely the time increment $t_n - t_{n-1}$, which is a tricky operation, and in (4.10) we sum a large number of approximations, and the question is then if the approximate sum, that is, the sum of approximations $f(t_{n-1})(t_n - t_{n-1})$ of $x(t_n) - x(t_{n-1})$, is a reasonable approximation of the "real" sum $x(1)$. Note that a sum of many small errors very well could result in an accumulated large error.

We have now gotten a first glimpse of Calculus. We repeat: the heart is the formula

$$x(t_n) \approx x(t_{n-1}) + f(t_{n-1})(t_n - t_{n-1})$$

or setting $f(t) = x'(t)$,

$$x(t_n) \approx x(t_{n-1}) + x'(t_{n-1})(t_n - t_{n-1}),$$

connecting increment in distance to velocity multiplied with increment in time. This formula contains both the definition of the integral reflecting (4.10) obtained after summation, and the definition of the derivative $x'(t)$ according to (4.6) obtained by dividing by $t_n - t_{n-1}$. Below we will uncover the surprising strength of these seemingly simple relations.

4.4 Generalization

We shall also meet the following generalization of (4.2)

$$x'(t) = f(x(t), t) \tag{4.12}$$

in which the function f on the right hand side depends not only on t but also on the unknown solution $x(t)$. The analog of formula (4.3) now may take the form

$$x(t_n) \approx x(t_{n-1}) + f(x(t_{n-1}), t_{n-1})(t_n - t_{n-1}), \tag{4.13}$$

or changing from t_{n-1} to t_n in the f-term and recalling (4.4),

$$x(t_n) \approx x(t_{n-1}) + f(x(t_n), t_n)(t_n - t_{n-1}), \tag{4.14}$$

where as above, $0 = t_0 < t_1 < \cdots < t_{n-1} < t_n \cdots < t_N = 1$ is a sequence of time instances.

Using (4.13), we may successively determine approximations of $x(t_n)$ for $n = 1, 2, ..., N$, assuming that $x(t_0)$ is a given initial value. If we use instead

(4.14), we obtain in each step an algebraic equation to determine $x(t_n)$ since the right hand side depends on $x(t_n)$.

In this way, solving the differential equation (4.12) approximately for $0 < t < 1$ is reduced to computing $x(t_n)$ for $n = 1, ..., N$, using the explicit formula (4.13) or solving the algebraic equation $X_n = X(t_{n-1}) + f(X_n, t_n)(t_n - t_{n-1})$ in the unknown X_n.

As a basic example, we will study the differential equation

$$x'(t) = x(t) \quad \text{for } t > 0, \tag{4.15}$$

corresponding to choosing $f(x(t), t) = x(t)$. In this case (4.13) takes the form

$$x(t_n) \approx x(t_{n-1}) + x(t_{n-1})(t_n - t_{n-1}) = (1 + (t_n - t_{n-1}))x(t_{n-1}).$$

With $(t_n - t_{n-1}) = \frac{1}{N}$ constant for $n = 1, ..., N$, we get the formula

$$x(t_n) \approx (1 + \frac{1}{N})x(t_{n-1}) \quad \text{for } n = 1, ..., N.$$

Repeating this formula, we get $x(t_n) \approx (1 + \frac{1}{N})(1 + \frac{1}{N})x(t_{n-2})$, and so on, which gives

$$x(1) \approx (1 + \frac{1}{N})^N x(0). \tag{4.16}$$

Later, we will see that there is indeed an exact solution of the equation $x'(t) = x(t)$ for $t > 0$ satisfying $x(0) = 1$, and we shall denote it $x(t) = \exp(t)$, and name it the *exponential function*. The formula (4.16) gives the following approximate formula for $\exp(1)$, where $\exp(1) = e$ is commonly referred to as the *base of the natural logarithm*:

$$e \approx (1 + \frac{1}{N})^N. \tag{4.17}$$

We give below values of $(1 + \frac{1}{N})^N$ for different N:

N	$(1 + \frac{1}{N})^N$
1	2
2	2.25
3	2.37
4	2.4414
5	2.4883
6	2.5216
7	2.5465
10	2.5937
20	2.6533
100	2.7048
1000	2.7169
10000	2.7181

The differential equation $x'(t) = x(t)$ for $t > 0$, models the evolution of for example a population of bacteria which grows at a rate $x'(t)$ equal to the given amount of bacteria $x(t)$ at each time instant t. After each one time unit such a population has multiplied with the factor $e \approx 2.72$.

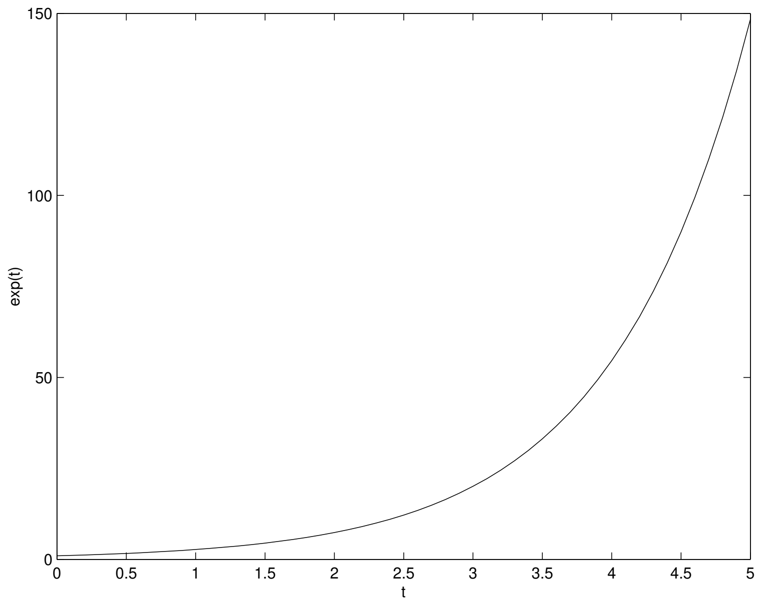


FIGURE 4.4. Graph of $\exp(t)$: Exponential growth.

4.5 Leibniz' teen-age dream

A form of Calculus was envisioned by Leibniz already as a teen-ager. Young Leibniz used to amuse himself with tables of the following form

n	1	2	3	4	5	6	7
n^2	1	4	9	16	25	36	49
	1	3	5	7	9	11	13
	1	2	2	2	2	2	2

or

n	1	2	3	4	5
n^3	1	8	27	64	125
	1	7	19	37	61
	1	6	12	18	24
	1	5	6	6	6

The pattern is that below each number, one puts the difference of that number and the number to its left. From this construction, it follows that any number in the table is equal to the sum of all the numbers in the next row below and to the left of the given number. For example for the squares n^2 in the first table we obtain the formula

$$n^2 = (2n-1) + (2(n-1)-1) + \cdots + (2\cdot 2 - 1) + (2 \cdot 1 - 1), \qquad (4.18)$$

which can also be written as

$$n^2 + n = 2(n + (n-1) + \cdots + 2 + 1) = 2\sum_{k=1}^{n} k. \qquad (4.19)$$

This corresponds to the area of the "triangular" domain in Fig. 4.5, where each term in the sum (the factor 2 included) corresponds to the area of one of the colons of squares.

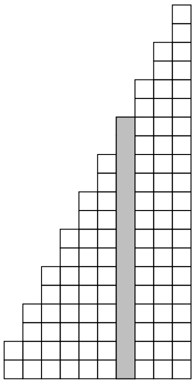


FIGURE 4.5.

The formula (4.19) is an analog of the formula

$$x^2 = 2\int_0^x y\,dy$$

with x corresponding to n, y to k, dy to 1, and $\sum_{k=1}^{n}$ to $\int_0^n$. Note that for n large the n-term in (4.19) is vanishing in comparison with n^2 in the sum $n^2 + n$.

By dividing by n^2, we can also write (4.18) as

$$1 = 2\sum_{k=1}^{n} \frac{k}{n}\frac{1}{n} - \frac{1}{n}, \qquad (4.20)$$

which is an analog of

$$1 = 2\int_0^1 y\,dy$$

with dy corresponding to $\frac{1}{n}$, y to $\frac{k}{n}$ and $\sum_{k=0}^{n}$ to $\int_0^1$. Note that the term $-\frac{1}{n}$ in (4.20) acts as a small error term that gets smaller with increasing n.

From the second table with n^3 we may similarly see that

$$n^3 = \sum_{k=1}^{n}(3k^2 - 3k + 1), \tag{4.21}$$

which is an analog of the formula

$$x^3 = \int_0^x 3y^2\, dy$$

with x corresponding to n, y to k and $dy = 1$.

By dividing by n^3, we can also write (4.21) as

$$1 = \sum_{k=0}^{n} 3(\frac{k}{n})^2 \frac{1}{n} - \frac{1}{n}\sum_{k=0}^{n} 3\frac{k}{n}\frac{1}{n} + \frac{1}{n^2},$$

which is an analog of

$$1 = \int_0^1 3y^2\, dy$$

with dy corresponding to $\frac{1}{n}$, y to $\frac{k}{n}$, and $\sum_{k=0}^{n}$ to $\int_0^1$. Again, the error terms that appear get smaller with increasing n.

Notice that repeated use of summation allows e.g. n^3 to be computed starting with the constant differences 6 and building the table from below.

Here is a description of Leibniz from Encyclopedia Brittannica: "Leibniz was a man of medium height with a stoop, broad-shouldered but bandy-legged, as capable of thinking for several days sitting in the same chair as of travelling the roads of Europe summer and winter. He was an indefatigable worker, a universal letter writer (he had more than 600 correspondents), a patriot and cosmopolitan, a great scientist, and one of the most powerful spirits of Western civilisation".

4.6 Summary

We may think of Calculus as the science of solving differential equations. With a similar sweeping statement, we may view Linear Algebra as the science of solving systems of algebraic equations. We may thus present the basic subjects of our study of Linear Algebra and Calculus in the form of the following two problems:

$$\text{Find } x \text{ such that } f(x) = 0 \quad \text{(algebraic equation)} \tag{4.22}$$

where $f(x)$ is a given function of x, and

$$\text{Find } x(t) \text{ such that } x'(t) = f(x(t), t) \quad \text{for } t \in (0,1],\ x(0) = 0, \quad \text{(differential equation)} \tag{4.23}$$

where $f(x,t)$ is a given function of x and t. Keeping this crude description in mind when following this book may help to organize the jungle of mathematical notation and techniques inherent to Linear Algebra and Calculus.

We shall largely take a *constructive* approach to the problem of solving equations, where we seek *algorithms* through which solutions may be determined or computed with more or less work. Algorithms are like recipes for finding solutions in a step by step manner. In the process of constructively solving equations we will need *numbers* of different kinds, such as *natural numbers*, *integers*, *rational numbers*. We will also need the concept of *real numbers*, *real variable*, *real-valued function*, *sequence of numbers*, *convergence*, *Cauchy sequence* and *Lipschitz continuous function*.

These concepts are supposed to be our humble servants and not terrorizing masters, as is often the case in mathematics education. To reach this position we will seek to demystify the concepts by using the constructive approach as much as possible. We will thus seek to look behind the curtain on the theater scene of mathematics, where often very impressive looking phenomena and tricks are presented by math teachers, and we will see that as students we can very well make these standard tricks ourselves, and in fact come up with some new tricks of our own which may even be better than the old ones.

4.7 Leibniz: Inventor of Calculus and Universal Genious

Gottfried Wilhelm von Leibniz (1646-1716) is maybe the most versatile scientist, mathematician and philosopher all times. Newton and Leibniz independently developed different formulations of Calculus; Leibniz notation and formalism quickly became popular and is the one used still today and which we will meet below. Leibniz boldly tackled the basic problem in Physics/Philosophy/Psychology of *Body and Soul* in his treatise *A New System of Nature and the Communication of Substances as well as the Union Existing between the Soul and the Body* from 1695. In this work Leibniz presented his theory of *Pre-established Harmony of Soul and Body*; In the related *Monadology* he describes the World as consisting of some kind of *elementary particles*in the form of *monads*, each of which with a blurred incomplete perception of the rest of the World and thus in possession of some kind of primitive soul. The modern variant of Monadology is

QuantuumMechanics, one of the most spectacular scientific achievements of the 20th century.

FIGURE 4.6. Leibniz, Inventor of Calculus: "Theoria cum praxis". "When I set myself to reflect on the Union of Soul with the Body, I seemed to be cast back again into the open sea. For I could find no way of explaining how the Body causes something to happen in the Soul, or vice versa.....Thus there remains only my hypothesis, that is to say *the way of the pre-established harmony*–pre-established, that is by a Divine anticipatory artifice, which is so formed each of theses substances from the beginning, that in merely following its own laws, which it received with its being, it is yet in accord with the other, just as if they mutually influenced one another, or as if, over and above his general concourse , God were for ever putting in his hands to set them right".

Chapter 4 Problems

4.1. Derive mathematical models of the form $y = f(x)$ connecting the displacement x with the force $y = f(x)$ for the following mechanical systems consisting of an elastic string in the first two cases and an elastic string coupled to an elastic spring in the third case. Find approximate models of the form $y = x^r$ with $r = 1, 3, \frac{1}{2}$.

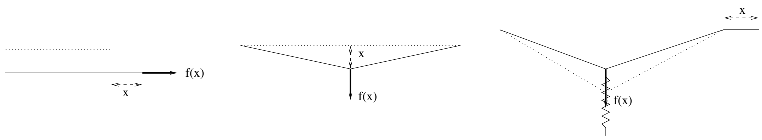


FIGURE 4.7.

4.2. Like Galileo solve the following differential equations: (a) $x'(t) = v$, (b) $x'(t) = at$, where v and a are constants. Interprets the results in a Pisa setting.

4.3. Consider the table

0	0	2	4	8	16	32	64
0	2	4	8	16	32	64	128
2	4	8	16	32	64	128	256

Do you see any connection to the exponential function?

4.4. The area of a disc of radius one is equal to π. Compute lower and upper bounds for π by comparing the area of the disc to the areas of inscribed and circumscribed polygons, see Fig. 4.8.

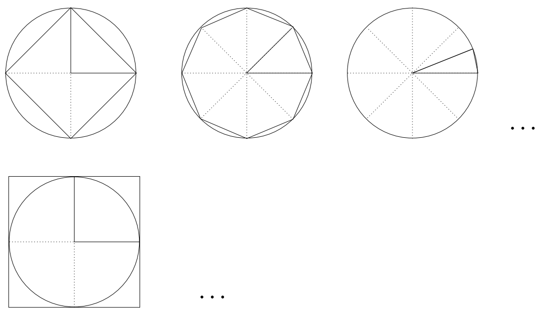


FIGURE 4.8.

4.5. Solve the Dinner Soup/Ice cream problem with the equation $2x = y + 1$ replaced by $x = 2y + 1$.

5
Natural Numbers and Integers

> "But", you might say, "none of this shakes my belief that 2 and 2 are 4". You are right, except in marginal case...and it is only in marginal cases that you are doubtful whether a certain animal is a dog or a certain length is less than a meter. Two must be two of something, and the proposition "2 and 2 are 4" is useless unless it can be applied. Two dogs and two dogs are certainly four dogs, but cases arrive in which you are doubtful whether two of them are dogs. "Well, at any rate there are four animals" you may say. But there are microorganisms concerning which it is doubtful whether they are animals or plants. "Well, then living organisms," you may say. But there are things of which it is doubtful whether they are living organisms or not. You will be driven into saying: "Two entities and two entities are four entities". When you have told me what you mean by "entity" I will resume the argument. (Russell)

5.1 Introduction

In this chapter, we recall how natural numbers and integers may be constructively defined, and how to prove the basic rules of computation we learn in school. The purpose is to give a quick example of developing a mathematical theory from a set of very basic facts. The idea is to give the reader the capability of explaining to her/his grandmother *why*, for example, 2 times 3 is equal to 3 times 2. Answering questions of this nature leads to a deeper understanding of the nature of integers and the rules for computing with integers, which goes beyond just accepting facts you learn in school as something given once and for all. An important aspect of this process is the very *questioning* of established facts that follows from posing the *why*, which may lead to new insight and new truths replacing the old ones.

5.2 The natural numbers

The *natural numbers* such as $1, 2, 3, 4, \ldots$, are familiar from our experience with *counting* where we repeatedly *add* 1 starting with 1. So $2 = 1 + 1$, $3 = 2+1 = 1+1+1, 4 = 3+1 = 1+1+1+1, 5 = 4+1 = 1+1+1+1+1$, and so on. Counting is a pervasive activity in human society: we count minutes waiting for the bus to come and the years of our life; the clerk counts change in the store, the teacher counts exam points, Robinson Crusoe counted the days by making cuts on a log. In each of these cases, the unit 1 represents something different; minutes and years, cents, exam points, days; but the process of counting is the same for all the cases. Children learn to count at an early age and may count to 10 by the age of say 3. Clever chimpanzees may also be taught to count to 10. The ability to count to 100 may be achieved by children of the age of 5.

The *sum* $n + m$ obtained by *adding* two natural numbers n and m, is the natural number resulting from adding 1 first n times and then m times. We refer to n and m as the *terms* of the sum $n+m$. The equality $2+3 = 5 = 3+2$ reflects that

$$(1 + 1) + (1 + 1 + 1) = 1 + 1 + 1 + 1 + 1 = (1 + 1 + 1) + (1 + 1),$$

which can be explained in words as observing that if we have 5 donuts in a box, then we can consume them by first eating 2 donuts and then 3 donuts or equally well by first eating 3 donuts and then 2 donuts. By the same argument we can prove the *commutative rule for addition*

$$m + n = n + m,$$

and the *associative rule for addition*

$$m + (n + p) = (m + n) + p,$$

where m, n, and p are natural numbers.

The *product* $m \times n = mn$ obtained by *multiplying* two natural numbers m and n, is the natural number resulting by adding n to itself m times. The numbers m and n of a product $m \times n$ are called *factors* of the product. The *commutative rule for multiplication*

$$m \times n = n \times m \tag{5.1}$$

expresses the fact that adding n to itself m times is equal to adding m to itself n times. This fact can be established by making a square array of dots with m rows and n columns and counting the total number of dots $m \times n$ in two ways: first by summing the m dots in each column and then summing over the n columns and second by summing the n dots in each row and then summing over the m rows, see Fig. 5.1. In a similar way we

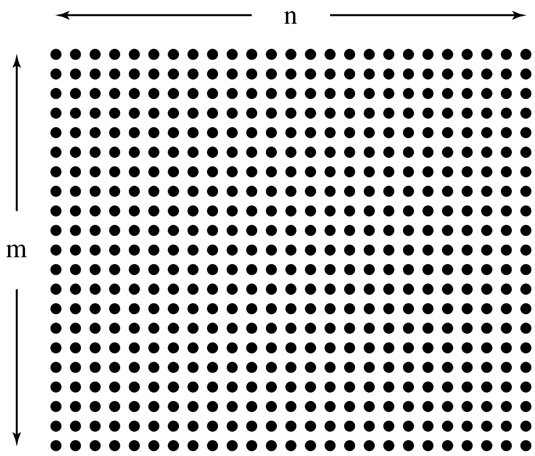


FIGURE 5.1. Illustration of the commutative rule for multiplication $m \times n = n \times m$. We get the same sum if first add up the dots by counting across the rows or down the columns.

can prove the *associative rule for multiplication*

$$m \times (n \times p) = (m \times n) \times p \qquad (5.2)$$

and the *distributive rule* combining addition and multiplication,

$$m \times (n + p) = m \times n + m \times p, \qquad (5.3)$$

for natural numbers m, n, and p. Note that here we use the *convention* that multiplications are carried out first, then summations, unless otherwise is indicated. For example, $2+3\times4$ means $2+(3\times4) = 24$, not $(2+3)\times4 = 20$. To overrule this convention we may use parentheses, as in $(2+3)\times4 = 5\times4$. From (5.3) (and (5.1)) we obtain the useful formula

$$(m + n)(p + q) = (m + n)p + (m + n)q = mp + np + mq + nq. \qquad (5.4)$$

We define $n^2 = n \times n$, $n^3 = n \times n \times n$, and more generally

$$n^p = \underbrace{n \times n \times \cdots \times n}_{(p \text{ factors})}$$

for natural numbers n and p, and refer to n^p as *n to the power p*, or the "p-th power of n". The basic properties

$$\begin{aligned}(n^p)^q &= n^{pq}\\ n^p \times n^q &= n^{p+q}\\ n^p \times m^p &= (nm)^p,\end{aligned}$$

follow directly from the definition, and from the associative and distributive laws of multiplication.

We also have a clear idea of ranking natural numbers according to size. We consider m to be larger than n, written as $m > n$, if we can obtain m by adding 1 repeatedly to n. The inequality relation satisfies its own set of rules including

$$\begin{gathered}m < n \text{ and } n < p \text{ implies } m < p\\ m < n \text{ implies } m + p < n + p\\ m < n \text{ implies } p \times m < p \times n\\ m < n \text{ and } p < q \text{ implies } m + p < n + q,\end{gathered}$$

which hold for natural numbers n, m, p, and q. Of course, $n > m$ is the same as $m < n$, and writing $m \leq n$ means that $m < n$ or $m = n$.

A way of representing the natural numbers is to use a horizontal line extending to the right with the marks 1, 2, 3, spaced at a unit distance consecutively, see Fig. 5.2. This is called the *natural number line*. The line serves like a ruler to keep the points lined up in ascending order to the right. We can interpret all of the arithmetic operations using the number

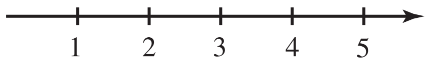


FIGURE 5.2. The natural number line.

line. For example, adding 1 to a natural number n means shifting one unit to the right from the position of n to that of $n + 1$, and likewise adding p means shifting p units to the right.

We can also extend the natural number line one unit to the left and mark that point by 0, which we refer to as *zero*. We can use 0 as a starting point from which we get to the point marked 1 by moving one unit to the right, We can interpret this operation as $0 + 1 = 1$, and generally we have

$$0 + n = n + 0 = n \tag{5.5}$$

for n a natural number. We further define $n \times 0 = 0 \times n = 0$ and $n^0 = 1$.

Representing natural numbers as sums of ones like $1 + 1 + 1 + 1 + 1$ or $1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1$, that is, as cuts on a log or as beads on

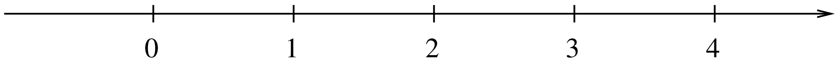


FIGURE 5.3. The extended natural number line, including 0.

a thread, quickly becomes impractical as the size of the number increases. To be able to express natural numbers of any size, it is convenient to use a *positional system*. In a positional system with *base 10* we use the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and express each natural number uniquely as a sum of terms of the form

$$d \times 10^p \tag{5.6}$$

where d is one of the digits $0, 1, 2, ..., 9$, and p is a natural number or 0. For example

$$4711 = 4 \times 10^3 + 7 \times 10^2 + 1 \times 10^1 + 1 \times 10^0.$$

We normally use the positional system with base 10, where the choice of base is of course connected to counting using our fingers.

One can use any natural number as the base in a positional system. The computer normally uses the *binary* system with base 2, where a natural number is expressed as a string of 0s and 1s. For example

$$1001 = 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0, \tag{5.7}$$

which equals the usual number 9. We will return to this topic below.

5.3 Is there a largest natural number?

The insight that counting always can be continued by adding 1 yet another time, that is the insight that if n is a natural number, then $n+1$ is a natural number, is an important step in the development of a child usually taken in early school years. Whatever natural number I would assign as the largest natural number, you could argue that the next natural number obtained by adding 1 is bigger, and I would probably have to admit that there cannot be a largest natural number. The line of natural numbers extends for ever to the right.

Of course, this is related to some kind of unlimited *thought experiment*. In reality, time or space could set limits. Eventually, Robinson's log would be filled with cuts, and a natural number with say 10^{50} digits would seem impossible to store in a computer since the number of atoms in the Universe is estimated to be of this order. The number of stars in the Universe is probably finite although we tend to think of this number as being without bound.

We may thus say that *in principle* there is no largest natural number, while *in practice* we will most likely never deal with natural numbers bigger than 10^{100}. Mathematicians are interested in principles and thus would like to first get across what is true in principle, and then at a later stage what may be true in practice. Other people may prefer to go to realities directly. Of course, principles may be very important and useful, but one should not forget that there is a difference between what is true in principle and what is really true.

The idea that, in principle, there cannot be a largest natural number, is intimately connected to the concept of *infinity*. We may say that there are *infinitely many* natural numbers, or that the *set of natural numbers is infinite*, in the sense that we can keep on counting or making cuts without ever stopping; there is always possible to make another cut and add 1 another time. With this view, the concept of infinity is not so difficult to grasp; it just means that we never come to an end. Infinitely many steps means a *potential* to take yet another step independent of the number of steps we have taken. There is no limit or bound. To have infinitely many donuts means that we can always take yet another donut *whenever we want independent of how many we have already eaten*. This potential seems more realistic (and pleasant) than actually eating infinitely many donuts.

5.4 The set $\mathbb{N}$ of all natural numbers

We may easily grasp the *set* $\{1, 2, 3, 4, 5\}$ of the first 5 natural numbers $1, 2, 3, 4, 5$. This may be done by writing down the numbers $1, 2, 3, 4$ and 5 on a piece of paper and viewing the numbers as constituting one entity, like a telephone number. We may even grasp the set $\{1, 2, ..., 100\}$ of the first 100 natural numbers $1, 2, 3, ..., 99, 100$ in the same way. We may also grasp individual very large numbers; for instance we might grasp the number 1 000 000 000 by imagining what we could buy for 1 000 000 000 dollars. We also feel quite comfortable with the principle of being able to add 1 to any given natural number. We could even agree to denote by $\mathbb{N}$ all the natural numbers that we potentially could reach by repeatedly adding 1.

We can think of $\mathbb{N}$ as the *set of possible natural numbers* and it is clear that this set is always under construction and can never actually be completed. It is like a highrise, where continuously new stores can be added on top without any limit set by city regulations or construction technique. We understand that $\mathbb{N}$ embodies a potential rather than an existing reality, as we discussed above.

The definition of $\mathbb{N}$ as the set of possible natural numbers is a bit vague because the term “possible” is a bit vague. We are used to the fact that what is possible for you may be impossible for me and vice versa. Whose “possible” should we use? With this perspective we leave the door a bit open to everyone to have his own idea of $\mathbb{N}$ depending on the meaning of “possible natural number” for each individual.

If we are not happy with this idea of $\mathbb{N}$ as "the set of *all possible* natural numbers", with its admitted vagueness, we may instead seek a definition of "the set of *all* natural numbers" which would be more universal. Of course any attempt to display this set by writing down all natural numbers on a piece of paper, would be rudely interrupted by reality. Deprived of this possibility, even in principle, it appears that we must seek guidelines from some Big Brother concerning the meaning of $\mathbb{N}$ as "the set of all natural numbers".

The idea of a universal Big Brother definition of difficult mathematical concepts connected to infinity one way or the other, like $\mathbb{N}$, grew strong during the late 19th century. The leader of this school was Cantor, who created a whole new theory dealing with infinite sets and infinite numbers. Cantor believed he could grasp the set of natural numbers as one completed entity and use this as a stepping stone to construct sets of even higher degrees of infinity. Cantors work had profound influence on the view of infinity in mathematics, but his theories about infinite sets were understood by few and used by even fewer. What remains of Cantors work today is a firm belief by a majority of mathematicians that the set of all natural numbers may be viewed as a uniquely defined completed entity which may be denoted by $\mathbb{N}$. A minority of mathematicians, the so-called constructivists led by Kronecker, have opposed Cantors ideas and would rather think of $\mathbb{N}$ somewhat more vaguely defined as the set of possible natural numbers, as we proposed above.

The net result appears to be that there is no consensus on the definition of $\mathbb{N}$. Whatever interpretation of $\mathbb{N}$ you prefer, and this is now open to your individual choice just as religion is, there will always remain some ambiguity to this notion. Of course, this reflects that we can give *names* to things that we cannot fully grasp, like *the world*, *soul*, *love*, *jazz music*, *ego*, *happiness* et cetera. We all have individual ideas of what these words mean.

Personally, we tend to favor the idea of using $\mathbb{N}$ to denote the "set of possible natural numbers". Admittedly this is a bit vague (but honest), and the vagueness does not appear to create any problems in our work.

5.5 Integers

If we associate addition by the natural number p as moving p units to the right on the natural number line, we can introduce the operation of *subtraction* by p as moving p units to the left. In the setting of donuts in a box, we can think of addition as putting donuts into the box and subtraction as taking them out. For example, if we have 12 donuts in the box and eat 7 of them, we know there will be 5 left. We originally got the 12 donuts by adding individual donuts into a box, and we may take away donuts, or subtract them, by taking them back out of the box. Mathematically, we write this as $12 - 7 = 5$ which is just another way of saying $5 + 7 = 12$.

We immediately run into a complication with subtraction that we did not meet with addition. While the sum $n + m$ of two natural numbers is always a natural number, the difference $n - m$ is a natural number only if $m < n$. Moving m units to the left from n will take us outside the natural number line if $m > n$. For example, the difference $12 - 15$ would arise if we wanted to take 15 donuts out of a box with 12 donuts. Similar situations arise frequently. If we want to buy a titanium bike frame for $2500, while we only have $1500 in the bank, we know we have to borrow $1000. This $1000 is a debt and does not represent a positive amount in our savings account, and thus does not correspond to a natural number.

To handle such situations, we *extend* the natural numbers $\{1, 2, 3, \cdots\}$ by adjoining the negative numbers $-1, -2, -3, \cdots$ together with 0. The result is the set of *integers*

$$\mathbb{Z} = \{\cdots, -3, -2, -1, 0, 1, 2, 3, \cdots\} = \{0, \pm 1, \pm 2, \pm 3, \cdots\}.$$

We say that $1, 2, 3, \cdots$, are the *positive integers* while $-1, -2, -3, \cdots$, are the *negative integers*. Graphically, we think of extending the natural number line to the left and then marking the point that is one unit distance to the left of 0 as -1, and so on, to get the *integer number line*, see Fig. 5.4.

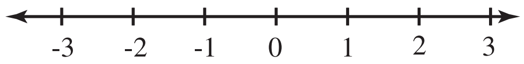


FIGURE 5.4. The integer number line.

We may define the sum $n + m$ of two integers n and m as the result of adding m to n as follows. If n and m are both natural numbers, or positive integers, then $n + m$ is obtained the usual way by starting at 0, moving n units to the right followed by m more units to the right. If n is positive and m is negative, then $n + m$ is obtained starting at 0, moving n units to the right, and then m units back to the left. Likewise if n is negative and m is positive, then we obtain $n + m$ by starting at 0, moving n units to the left and then m units to the right. Finally, if both n and m are negative, then we obtain $n + m$ by starting at 0, moving n units to the left and then m more units to the left. Adding 0, we move neither right nor left, and thus $n + 0 = n$ for all integers n. We have now extended the operation of addition from the natural numbers to the integers.

Next, to define the operation of *subtraction*, we first agree to denote by $-n$ the integer with the opposite sign to the integer n. We then have for any integer n that $-(-n) = n$, reflecting that taking the opposite sign twice gives the original sign, and $n + (-n) = (-n) + n = 0$, reflecting that moving n units back and forth starting at 0 will end up at 0. We now define $n - m = -m + n = n + (-m)$, which we refer to as *subtracting* m from n. We see that subtracting m from n is the same as adding $-m$ to n.

Finally, we need to extend multiplication to integers. To see how to do this, we seek guidance by formally multiplying the equality $n + (-n) = 0$, where n a natural number, by the natural number m. We then obtain $m \times n + m \times (-n) = 0$, which suggest that $m \times (-n) = -(m \times n)$, since $m \times n + (-(m \times n)) = 0$. We are thus led to define $m \times (-n) = -(m \times n)$ for positive integers m and n, and likewise $(-n) \times m = -(n \times m)$. Note that by this definition, $-n \times m$ may be interpreted both as $(-n) \times m$ and as $-(n \times m)$. In particular we have that $(-1) \times n = -n$ for n a positive integer. Finally, to see how to define $(-n) \times (-m)$ for n and m positive integers, we multiply the equalities $n + (-n) = 0$ and $m + (-m) = 0$ to get formally $n \times m + n \times (-m) + (-n) \times m + (-n) \times (-m) = 0$, which indicates that $-n \times m + (-n) \times (-m) = 0$, that is $(-n) \times (-m) = n \times m$, which we now take as a definition of the product of two negative numbers $(-n)$ and $(-m)$. In particular we have $(-1) \times (-1) = 1$. We have now defined the product of two arbitrary integers (of course we set $n \times 0 = 0 \times n = 0$ for any integer n).

To sum up, we have defined the operations of addition and multiplication of integers and we can now verify all the familiar rules for computing with integers including the commutative, associative and distributive rules stated above for natural numbers.

Note that we may say that we have *constructed* the negative integers $\{-1, -2, ...\}$ from the given natural numbers $\{1, 2, ...\}$ through a process of reflection around 0, where each natural number n gets its mirror image $-n$. We thus may say that we *construct* the integer line from the natural number line through a process of reflection around 0. Kronecker said that the natural numbers were given by God and that all other numbers, like the negative integers, are invented or constructed by man.

Another way to define or construct $-n$ for a natural number n is to think of $-n$ as the solution $x = -n$ of the equation $n + x = 0$ since $n + (-n) = 0$, or equally well as the solution of $x + n = 0$ since $(-n) + n = 0$. This idea is easily extended from n to $-n$, i.e. to the negative integers, by considering $-(-n)$ to be the solution of $x + (-n) = 0$. Since $n + (-n) = 0$, we conclude the familiar formula $-(-n) = n$. To sum up, we may view $-n$ to be the solution of the equation $x + n = 0$ for any integer n.

We further extend the ordering of the natural numbers to all of $\mathbb{Z}$ by defining $m < n$ if m is to the left of n on the integer line, that is, if m is negative and n positive, or zero, or if also n is negative but $-m > -n$. This ordering is a little bit confusing, because we like to think of for example -1000 as a lot bigger number than -10. Yet we write $-1000 < -10$ saying that -1000 is smaller than -10. What we need is a measure of the *size* of a number, disregarding its sign. This will be the topic next.

5.6 Absolute value and the distance between numbers

As just indicated, it is convenient to be able to discuss the *size* of numbers independent of the sign of the number. For this purpose we define the ***absolute value*** $|p|$ of the number p by

$$|p| = \begin{cases} p, & p \geq 0 \\ -p, & p < 0. \end{cases}$$

For example, $|3| = 3$ and $|-3| = 3$. Thus $|p|$ measures the *size* of the number p, disregarding its sign, as desired. For example $|-1000| > |-10|$.

Often we are interested in the difference between two numbers p and q, but are concerned primarily with the *size* of the difference and care less about its sign, that is we are interested in $|p-q|$ corresponding to the *distance* between the two numbers on the number line.

For example suppose we have to buy a piece of molding for a doorway and when using a tape measure we position one side of the doorframe at 2 inches and the opposite side at 32 inches. We would not go to the store and ask the person for a piece of molding that begins at 2 inches and ends at 32 inches. Instead, we would only tell the clerk that we need $32-2=30$ inches. In this case, 30 is the distance between 32 and 2. We define the ***distance*** between two integers p and q as $|p-q|$.

By using the absolute value, we insure that the distance between p and q is the same as the distance between q and p. For example, $|5-2| = |2-5|$.

In this book, we will be dealing with inequalities combined with the absolute value frequently. We give an example close to every student's heart.

EXAMPLE 5.1. Suppose the scores on an exam that are within 5 of 79 out of 100 get a grade of B and we want to write down the list of scores that get a B. This includes all scores x that are a distance of at most 5 from 79, which can be written

$$|x - 79| \leq 5. \tag{5.8}$$

There are two possible cases: $x < 79$ and $x \geq 79$. If $x \geq 79$ then $|x-79| = x-79$ and (5.8) becomes $x - 79 \leq 5$ or $x \leq 84$. If $x < 79$ then $|x-79| = -(x-79)$ and (5.8) means that $-(x-79) \leq 5$ or $(x-79) \geq -5$ or $x \geq 74$. Combining these results we have $79 \leq x \leq 84$ as one possibility or $74 \leq x < 79$ as another possibility, or in other words, $74 \leq x \leq 84$.

In general if $|x| < b$, then we have the two possibilities $-b < x < 0$ or $0 \leq x < b$ which means that $-b < x < b$. We can actually solve both cases at one time.

EXAMPLE 5.2. $|x - 79| \leq 5$ means that

$$\begin{array}{rcccl} -5 & \leq & x - 79 & \leq & 5 \\ 74 & \leq & x & \leq & 84 \end{array}$$

To solve $|4 - x| \leq 18$, we write

$$\begin{array}{rcccl} -18 & \leq & 4 - x & \leq & 18 \\ 18 & \geq & x - 4 & \geq & -18 \text{ \textit{(Note the changes!)}} \\ 22 & \geq & x & \geq & -14 \end{array}$$

EXAMPLE 5.3. To solve the following inequality in x:

$$|x - 79| \geq 5. \tag{5.9}$$

we first assume that $x \geq 79$, in which case (5.9) becomes $x - 79 \geq 5$ or $x \geq 84$. Next, if $x \leq 79$ then (5.9) becomes $-(x - 79) \geq 5$ or $(x - 79) \leq -5$ or $x \leq -74$. The answer is thus all x with $x \geq 84$ or $x \leq -74$.

Finally we recall that multiplying an inequality by a negative number like (-1) reverses the inequality:

$$m < n \text{ implies } -m > -n.$$

5.7 Division with remainder

We define *division with remainder* of a natural number n by another natural number m, as the process of computing nonnegative integers p and $r < m$ such that $n = pm+r$. The existence of unique p and r follows by considering the sequence of natural numbers m, $2m$, $3m$,..., and noting that there must be a unique p such that $pm \leq n < (p+1)m$, see Fig. 5.5.

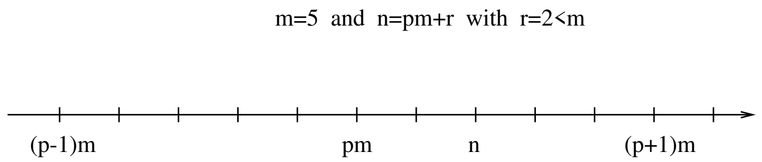


FIGURE 5.5. Illustration of $pm \leq n < (p+1)m$.

Setting $r = n - pm$, we obtain the desired representation $n = pm + r$ with $0 \leq r < m$. We call r the *remainder* in division of n by m. When the remainder r is zero, then we obtain a *factorization* $n = pm$ of n as a product of the *factors* p and m.

We can find the proper p in division with remainder of n by m by repeated subtraction of m. For example, if $n = 63$ and $m = 15$, then we may write

$$63 = 15 + 48$$
$$63 = 15 + 15 + 34 = 2 \times 15 + 33$$
$$63 = 3 \times 15 + 18$$
$$63 = 4 \times 15 + 3,$$

and thus find that in this case $p = 4$ and $r = 3$.

A more systematic procedure for division with remainder is the *long division* algorithm, which is taught in school. We give two examples ($63 = 4 \times 15 + 3$ again, and $2418610 = 19044 \times 127 + 22$) in Fig. 5.5.

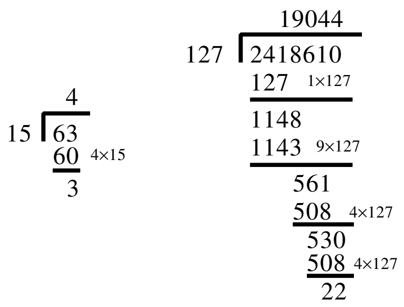


FIGURE 5.6. Two examples of long division.

5.8 Factorization into prime factors

A *factor* of a natural number n is a natural number m that divides into n without leaving a remainder, that is, $n = pm$ for some natural number p. For example, 2 and 3 are both factors of 6. A natural number n always has factors 1 and n since $1 \times n = n$. A natural number n is called a *prime number* if the only factors of n are 1 and n. The first few prime numbers (excluding 1 since such factors are not of much interest) are $\{2, 3, 5, 7, 11, \cdots\}$. The only even prime number is 2. Suppose that we take the natural number n and try to find two factors $n = pq$. Now there are two possibilities: either the only two factors are 1 and n, i.e. n is prime, or we find two factors p and q, neither of which are 1 or n. By the way, it is easy to write a program to search for all the factors of a given natural number n by systematically dividing by all the natural numbers up to n. Now in the second case, both p and q must be less than n. In fact $p \leq n/2$ and $q \leq n/2$ since the smallest possible factor not equal to 1 is 2. Now we repeat by factoring p and q separately. In each case, we either find the number is prime or we factor it into a product of smaller natural numbers. Then we continue with the smaller factors. Eventually this process must stop since n is finite and the factors at any stage are no larger than half the size of the factors of the previous stage. When the process has stopped, we have *factored* n into a product of prime numbers. This factorization is unique except for order. One consequence of the factorization into prime numbers is the following fact. Suppose that we know that 2 is a factor of n. If $n = pq$ is any factorization of n, it follows that at least one of the factors p and q must have a factor of 2. The same is true for prime number factors 3, 5, 7 etc., that is for any prime number factor.

5.9 Computer representation of integers

Since we will be using the computer throughout this course, we have to point out some properties of computer arithmetic. We are distinguishing arithmetic carried out on a computer from the "theoretical" arithmetic we learn about in school.

The fundamental issue that arises when using a computer stems from the physical limitation on memory. A computer must store numbers on a physical device which cannot be "infinite". Hence, *a computer can only represent a finite number of numbers.* Every computer language has a finite limit on the numbers it can represent. It is quite common for a computer language to have *INTEGER* and *LONG INTEGER* types of variables, where an INTEGER variable is an integer in the range of $\{-32768, -32767, ..., 32767\}$, which are the numbers that take two bytes of storage, and a long integer variable is an integer in the range $\{-2147483648,\ -2147483647,\ ...,\ 2147483647\}$, which are the integers requiring four bytes of storage (where a "byte" of memory consists of 8 "bit-cells", each capable of storing either a zero or a one). This can have some serious consequences, as anyone who programs a loop using an integer index that goes above the appropriate limit finds out. In particular, we cannot check whether some fact is true for all integers using a computer to test each case.

Chapter 5 Problems

5.1. Identify five ways in your life in which you count and the unit "1" for each case.

5.2. Use the natural number line representation to interpret and verify the equalities: (a) $x + y = y + x$ and (b) $x + (y + z) = (x + y) + z$: that hold for any natural numbers x, y, and z:

5.3. Use (two and three dimensional) arrays of dots to interpret and verify a) the distributive rule for multiplication $m \times (n + p) = m \times n + m \times p$ and b) the associative rule $(m \times n) \times p = m \times (n \times p)$..

5.4. Use the definition of n^p for natural numbers n and p to verify that (a) $(n^p)^q = n^{pq}$ and (b) $n^p \times n^q = n^{p+q}$ for natural numbers n, p, q.

5.5. Prove that $m \times n = 0$ if and only if $m = 0$ or $n = 0$, for integers m and n. What does *or* mean here? Prove that for $p \neq 0$, $p \times m = p \times n$ if and only if $m = n$. What can be said if $p = 0$?

5.6. Verify using (5.4) that for integers n and m,

$$\begin{aligned}(n+m)^2 &= n^2 + 2nm + m^2 \\ (n+m)^3 &= n^3 + 3n^2m + 3nm^2 + m^3 \\ (n+m)(n-m) &= n^2 - m^2.\end{aligned} \tag{5.10}$$

5.7. Use the integer number line to illustrate the four possible cases in the definition of $n + m$ for integers n and m.

5.8. Divide (a) 102 by 18, (b) -4301 by 63, and (c) 650912 by 309 using long division.

5.9. (a) Find all the natural numbers that divide into 40 with zero remainder. (b) Do the same for 80.

5.10. *(Abstract)* Use long division to show that

$$\frac{a^3 + 3a^2b + 3ab^2 + b^3}{a + b} = a^2 + 2ab + b^2.$$

5.11. (a) Write a *MATLAB*© routine that tests a given natural number n to see if it is prime. Hint: systematically divide n by the smaller natural numbers from 2 to $n/2$ to check whether there are factors. Explain why it suffices to check up to $n/2$. (b) Use this routine to write a *MATLAB*© routine that finds all the prime numbers less than a given number n. (c) List all the prime numbers less than 1000.

5.12. Factor the following integers into a product of prime numbers; (a) 60, (b) 96, (c) 112, (d) 129.

5.13. Find two natural numbers p and q such that pq contains a factor of 4 but neither p nor q contains a factor of 4. This means that the fact that some natural number m is factor of a product $n = pq$ does not imply that m must be a factor of either p or q. Why doesn't this contradict the fact that if pq contains a factor of 2 then at least one of p or q contains a factor of 2?

5.14. Pick out the *invalid* rules from the following list

$$a < b \text{ implies } a - c < b - c$$
$$(a + b)^2 = a^2 + b^2$$
$$\big(c(a + b)\big)^2 = c^2(a + b)^2$$
$$ac < bc \text{ implies } a < b$$
$$a - b < c \text{ implies } a < c + b$$
$$a + bc = (a + b)c$$

In each case, find numbers that show the rule is invalid.

5.15. Solve the following inequalities:

(a) $|2x - 18| \leq 22$ (b) $|14 - x| < 6$

(c) $|x - 6| > 19$ (d) $|2 - x| \geq 1$

5.16. Verify that the following is true for arbitrary integers a, b and c: (a) $|a^2| = a^2$ (b) $|a|^2 = a^2$ (c) $|ab| = |a|\,|b|$ (d) $|a + b| \leq |a| + |b|$ (e) $|a - b| \leq |a| + |b|$ (f) $|a + b - c| \leq |a| + |b| + |c|$ (g) $|a| \leq |a - b| + |b|$ (h) $||a| - |b|| \leq |a - b|$

5.17. Show that the inequalities (e)-(h) of Problem 5.17 follow once you have (d) and the fact that $|a| = |-a|$ for any integer a.

5.18. Write a little program in the computer language of your choice that finds the largest integer that the language can represent. Hint: usually one of two things happen if you try to set an integer variable to a value that is too large: either you get an error message or the computer gives the variable a negative value.

6
Mathematical induction

> There is a tradition of opposition between adherents of induction and deduction. In my view it would be just as sensible for the two ends of a worm to quarrel. (Whitehead)

6.1 Induction

Carl Friedrich Gauss (1777-1885), sometimes called the Prince of Mathematics, is one of the greatest mathematicians all times. In addition to an incredible ability to compute (especially important in the 1800s) and an unsurpassed talent for mathematical proof, Gauss had an inventive imagination and a restless interest in nature and he made important discoveries in a staggering range of pure and applied mathematics. He was also a pioneer in the constructionist sense, digging deeply into many of the accepted mathematical truths of his time in order to really understand what everyone "knew" had to be true. Perhaps the only really unfortunate side to Gauss is that he wrote about his work only very sparingly and many mathematicians that followed him were doomed to reinvent things that he already knew.

There is a story about Gauss in school at the age of ten which goes as follows. His old-fashioned arithmetic teacher would like to show off to his students by asking them to add a large number of sequential numbers by hand, something the teacher knew (from a book) could be done quickly and accurately by using the following neat formula:

$$1 + 2 + 3 + \cdots + (n-1) + n = \frac{n(n+1)}{2}. \tag{6.1}$$

Note that the "$\cdots$" indicate that we add all the natural numbers between 1 and n. Using the formula makes it possible to replace the $n-1$ additions on the left by a multiplication and a division, which is a considerable reduction in work, especially if you are using a piece of chalk and a slate to do the sums.

The teacher posed the problem of computing the sum $1 + 2 + \cdots + 99$ to the class, and almost immediately Gauss came up and laid his slate down on the desk with the correct answer (4950), while the rest of the class still were in the beginning of a long struggle. How did young Gauss manage to compute the sum so quickly? Did he already know the nice formula (6.1)? Of course not, but he immediately derived it using the following clever argument: To sum $1 + 2 + \cdots + 99$, group the numbers two by two as follows:

$$\begin{aligned} &1 + \cdots + 99 \\ &\quad = (1 + 99) + (2 + 98) + (3 + 97) + \cdots (49 + 51) + 50 \\ &\quad = 49 \times 100 + 50 = 49 \times 2 \times 50 + 50 = 99 \times 50 \end{aligned}$$

which agrees with the formula (6.1) with $n = 99$. One can use this type of argument to prove the validity of (6.1) for any natural number n.

A modern teacher without a need to show off, could state the formula (6.1) and ask the students to use it with $n = 99$, for example, and then go home and play the trick with their parents.

Let's now look at the problem of verifying that the formula (6.1) is true for any natural number n, once this formula is given to us. Suppose, we are not as clever as Gauss and don't find the nice way of grouping the numbers two by two indicated above. We are thus asked just to *check* if a given formula is correct, and not to first *find* the formula itself and then prove its validity. This is like in a multiple choice test, where we may be asked if King Gustav Adolf II of Sweden died 1632 or 1932 or not at all, as opposed to a direct question what year this king died. Everyone knows that a multiple choice questions may be easier to answer than a direct question.

We are thus asked to prove that the formula (6.1) holds for *any* natural number n. It is easy enough to verify that it is true for $n = 1$: $1 = 1 \times 2/2$: for $n = 2$: $1 + 2 = 3 = 2 \times 3/2$: and for $n = 3$: $1 + 2 + 3 = 6 = 3 \times 4/2$. Checking the validity in this way for any natural number, one at a time, up to say $n = 1000$, would be very tiring. Of course we could try to get help from a computer, but the computer would also get stuck if n is very large. We also know in the back of our minds that no matter how many natural numbers n we check, there will always be natural numbers left which we have not yet checked.

Is there some other way of checking the validity of (6.1) for any natural number n? Yes, we could try the principle of *mathematical induction*, based on the idea of showing that if the formula is valid for n, then it is *automatically* valid also for $n+1$.

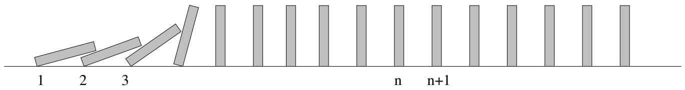


FIGURE 6.1. The principle of induction: If n falls then $n+1$ will fall, so if 1 falls they all fall!

The first step is then to check that the formula is valid for $n = 1$. We already took this easy step. The remaining step, which is called the *inductive step*, is to show that if the formula holds for a certain natural number, then it also holds for the next natural number. The principle of mathematical induction now states that the formula must be true for any natural number n. You are probably ready to accept this principle on intuitive grounds: we know that the formula holds for $n = 1$. By the inductive step, it then holds for the next number, that is for $n = 2$, and thus for $n = 3$ again by the induction step, and then for $n = 4$, and so on. Since we will eventually reach any natural number this way, we may be pretty sure that the formula holds for any natural number. Of course the principle of mathematical induction is based on the conviction that we will eventually reach any natural number if we start with 1 and then add 1 sufficiently many times.

Let us now take the induction step in the attempted proof of (6.1). We thus *assume* that the formula (6.1) is valid for $n = m - 1$, where $m \geq 2$ is a natural number. In other words, we assume that

$$1 + 2 + 3 + \cdots + m - 1 = \frac{(m-1)m}{2}. \tag{6.2}$$

We now want to *prove* that the formula holds for the next natural number $n = m$. To do this for (6.1), we add m to both sides of (6.2) to get

$$\begin{aligned} 1 + 2 + 3 + \cdots + m - 1 + m &= \frac{(m-1)m}{2} + m \\ &= \frac{m^2 - m}{2} + \frac{2m}{2} = \frac{m^2 - m + 2m}{2} \\ &= \frac{m(m+1)}{2} \end{aligned}$$

which shows the validity of the formula for $n = m$. We have thus verified that if (6.1) is true for a certain natural number $n = m - 1$ then it is true for the next natural number $n = m$, that is we have verified the inductive step. Repeating the inductive step, we see that (6.1) holds for any natural number n.

We may phrase the proof of the inductive step alternatively as follows, where we don't bother to introduce the natural number m. In the reformulation we assume that (6.1) holds with n replaced by $n-1$, that is, we assume that

$$1+2+3+\cdots+n-1=\frac{(n-1)n}{2}.$$

Adding n to both sides, we get

$$1+\cdots+n-1+n=\frac{(n-1)n}{2}+n=\frac{n(n+1)}{2}$$

which is (6.1) as desired. We may thus take the inductive step by proving (6.1) assuming the validity of (6.1) with n replaced by $n-1$.

Formulas like (6.1) have a close connection to the basic integration formulas of Calculus we meet later and they also occur, for example, when computing compound interest on a savings account or adding up populations of animals. The formulas are thus useful not only for impressing people.

Many students may find that verifying a property like (6.1) for specific n like 1 or 2 or 100 is not so difficult to do. But the general inductive step: assuming the formula is true for some given natural number and showing that it is then true for the next natural number, causes some uneasiness because the given number is not specified concretely. Don't let this feeling of strangeness stop you from trying out the problems: like much of mathematics, actually doing the problems is not as bad as the anticipation of having to do the problems (like going to the dentist). Working out some induction problems is good practice for getting used to some of the arguments that we encounter later.

The method of induction may be useful for showing the validity of a given formula, but first you have to *find* the formula in some way, which may require some good intuition, trial and error, or some other insight (like the clever idea of Gauss). Induction may thus help you because it gives you some kind of method of approach (assume the validity for some natural number and then prove validity for the next), but you must have a good guess or conjecture to start from.

You could try to come up with the formula (6.1) by some trial and error, once you have the courage to start looking for a formula. You could argue roughly that the average size of the numbers 1 to n is $n/2$, and since there are n numbers to add, their sum should be something like $n\frac{n}{2}$, which is pretty close to the correct $(n+1)\frac{n}{2}$. You don't need to be Gauss to see this.

We now give three additional examples illustrating the use of mathematical induction.

EXAMPLE 6.1. First we show a formula for the *geometric sum* with n terms for the fixed natural number $p > 1$,

$$1 + p + p^2 + p^3 + \cdots + p^n = \frac{1 - p^{n+1}}{1 - p} \tag{6.3}$$

which holds for any natural number n. Note that we think of p as being fixed and the induction is on n. The formula (6.3) holds for $n = 1$, since

$$1 + p = \frac{(1-p)(1+p)}{1-p} = \frac{1-p^2}{1-p},$$

where we use the formula $a^2 - b^2 = (a-b)(a+b)$. Assuming it is true with n replaced by $n - 1$, we have

$$1 + p + p^2 + p^3 + \cdots + p^{n-1} = \frac{1-p^n}{1-p}.$$

We add p^n to both sides to get

$$\begin{aligned} 1 + p + p^2 + p^3 + \cdots + p^{n-1} + p^n &= \frac{1-p^n}{1-p} + p^n \\ &= \frac{1-p^n}{1-p} + \frac{p^n(1-p)}{1-p} \\ &= \frac{1-p^{n+1}}{1-p} \end{aligned}$$

which shows the inductive step.

Of course, a much simpler way to verify (6.3) is to just note that $(1 - p)(1 + p + p^2 + p^3 + \cdots + p^n) = 1 + p + p^2 + p^3 + \cdots + p^n - p - p^2 - p^3 - \cdots - p^n - p^{n+1} = 1 - p^{n+1}$, and then simply divide by $1 - p$.

EXAMPLE 6.2. Induction can also be used to show properties that do not involve sums. For example, we show an inequality that is useful. For any fixed natural number p,

$$(1 + p)^n \geq 1 + np \tag{6.4}$$

for any natural number n. The inequality (6.4) is certainly valid for $n = 1$, since $(1 + p)^1 = 1 + 1 \times p$. Now assume it holds for $n - 1$,

$$(1 + p)^{n-1} \geq 1 + (n-1)p.$$

We multiply both sides by the positive number $1 + p$,

$$\begin{aligned} (1+p)^n = (1+p)^{n-1}(1+p) &\geq (1 + (n-1)p)(1+p) \\ &\geq 1 + (n-1)p + p + (n-1)p^2 \\ &\geq 1 + np + (n-1)p^2. \end{aligned}$$

Since $(n-1)p^2$ is nonnegative, we can take it away from the right-hand side and then obtain (6.4).

6.2 Changes in a population of insects

Induction is also often used to derive models. We now present an example involving the growth in a population of insects. We consider a simplified situation of a population of an insect in which the adults breed during the first summer they are alive then die before the following summer. We want to figure out how the population of these insects changes year by year. This is an important issue for instance if the insects carry disease or consume farm crops. In general, there are many factors that affect the rate of reproduction; the food supply, the weather, pesticides, and even the population itself. But in first phase of the modelling, we simplify all of this by assuming that the number of offspring produced each breeding season is simply *proportional to* the number of insects alive during that season. Experimentally this is often a valid assumption if the population is not too large.

Because we are describing populations of the insects during different years, we need to introduce a notation that makes it easy to associate variable names with different years. We use the index notation to do this. We let P_0 denote the current or *initial* population and $P_1, P_2, \cdots, P_n, \cdots$ denote the populations during subsequent years number $1, 2, \cdots, n, \cdots$ respectively. The *index* or *subscript* on P_n is a convenient way to denote the year. Following our assumption, we know that P_n is always proportional to P_{n-1}. Our modelling assumption is that the population P_n, after any number of years n, is *proportional to*, that is a fixed multiple of, the population P_{n-1} the year before. Using R to denote the constant of proportionality, we thus have

$$P_n = RP_{n-1}. \tag{6.5}$$

Assuming the initial population P_0 is known, the problem is to figure out when the population reaches a specific level M. In other words, find the first n such that $P_n \geq M$.

In order to do this, we want to find a formula expressing the dependence of P_n on n. We can do this by using induction on (6.5). Since (6.5) also holds for $n-1$, i.e. $P_{n-1} = RP_{n-2}$. Substituting, we find

$$P_n = RP_{n-1} = R(RP_{n-2}) = R^2 P_{n-2}.$$

Now we substitute for $P_{n-2} = RP_{n-3}$, $P_{n-3} = RP_{n-4}$, and so on. After $n-2$ more substitutions, we find

$$P_n = R^n P_0. \tag{6.6}$$

Since R and P_0 are known, this gives an explicit formula for P_n in terms of n. Note that the way we use induction in this example might seem different than the previous examples. But the difference is only superficial. To make the induction argument look the same as for the previous examples, we can assume that (6.6) holds for $n-1$ and then use (6.5) to show that it therefore holds for n.

Returning to the question of finding n such that $P_n \geq M$, the model problem is to find n such that

$$R^n \geq M/P_0. \tag{6.7}$$

As long as $R > 1$, R^n eventually grows large enough to do this. For example, if $R = 2$, then P_n grows quickly with n. If $P_0 = 1\,000$, then $P_1 = 2\,000$, $P_5 = 32\,000$, and $P_{10} = 1\,024\,000$.

Chapter 6 Problems

6.1. Prove the following formulas:

$$\text{(a)} \qquad 1^2 + 2^2 + 3^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6} \tag{6.8}$$

and

$$\text{(b)} \qquad 1^3 + 2^3 + 3^3 + \cdots + n^3 = \left(\frac{n(n+1)}{2}\right)^2 : \tag{6.9}$$

hold for all natural numbers n by using induction.

6.2. Using induction, show the following formula holds for all natural numbers n,

$$\frac{1}{1\times 2} + \frac{1}{2\times 3} + \frac{1}{3\times 4} + \cdots + \frac{1}{n(n+1)} = \frac{1}{n+1}.$$

6.3. Using induction, show the following inequalities hold for all natural numbers n:

$$\text{(a) } 3n^2 \geq 2n+1 \qquad \text{(b) } 4^n \geq n^2$$

6.4. The problem is to model the population of a species of insects that has a single breeding season during the summer. The adults breed during the first summer they are alive then die before the following summer. Assuming that the number of offspring born each breeding season is proportional to the square of the number of adults, express the population of the insects as a function of the year.

6.5. The problem is to model the population of a species of insects that has a single breeding season during the summer. The adults breed during the first summer they are alive then die before the following summer and moreover the adults kill and eat some of their offspring. Assuming that the number of offspring born each breeding season is proportional to the number of adults and that a number of offspring are killed by the adults is proportional to the square of the number of adults, derive an equation relating the population of the insects in one year to the population in the previous year.

6.6. *(Harder)* The problem is to model the population of a species of insects that has a single breeding season during the summer. The adults breed during the first and second summers they are alive then die before their third summer. Assuming that the number of offspring born each breeding season is proportional to the number of adults that are alive, derive an equation relating the population of the insects in any year past the first to the population in the previous two years.

6.7. Derive the formula (6.1) by verifying the sum

$$\begin{array}{rrrrrrrrrr} & 1 & + & 2 & + & \cdots & + & n-1 & + & n \\ + & n & + & n-1 & + & \cdots & + & 2 & + & 1 \\ \hline & n+1 & + & n+1 & + & \cdots & + & n+1 & + & n+1 \end{array}$$

and showing that this means that $2(1+2+\cdots+n) = n \times (n+1)$.

6.8. *(Harder)* Prove the formula for the geometric sum (6.3) holds by using induction on long division on the expression

$$\frac{p^{n+1}-1}{p-1}.$$

FIGURE 6.2. Gauss 1832

7
Rational numbers

The chief aim of all investigations of the external world should be to discover the rational order and harmony which has been imposed on it by God and which He revealed to us in the language of mathematics. (Kepler)

7.1 Introduction

We learn in school that a *rational number* r is a number of the form $r = \frac{p}{q} = p/q$, where p and q are integers with $q \neq 0$. Such numbers are also refereed to as fractions or ratios or quotients. We call p the *numerator* and q the *denominator* of the fraction or ratio. We know that $\frac{p}{1} = p$, and thus the rational numbers include the integers. A basic motivation for the invention of rational numbers is that with them we can solve equations of the form

$$qx = p$$

with p and $q \neq 0$ integers. The solution is $x = \frac{p}{q}$. In the Dinner Soup model we met the equation $15x = 10$ of this form with solution $x = \frac{10}{15} = \frac{2}{3}$. Clearly, we could not solve the equation $15x = 10$ if x was restricted to be a natural number, so you and your roommate should be happy to have access to the rational numbers.

If the natural number m is a factor of the natural number n so that $n = pm$ with p a natural number, then $p = \frac{n}{m}$, in which case thus $\frac{n}{m}$ is a natural number. If division of n by m leaves a non-zero remainder r, so that $n = pm + r$ with $0 < r < m$, then $\frac{n}{m} = p + \frac{r}{m}$, which is not a natural number.

7.2 How to construct the rational numbers

Suppose now that your roommate has an unusual background and has never heard about rational numbers, but fortunately is very familiar with integers and is more than willing to learn new things. How could you quickly explain to her/him what rational numbers *are* and how to compute with them? In other words, how could you convey how to *construct* rational numbers from integers, and how to add, subtract, multiply and divide rational numbers? One possibility would be to simply say that $x = \frac{p}{q}$ is "that thing" which solves the equation $qx = p$, with p and $q \neq 0$ integers. For example, a quick way to convey the meaning of $\frac{1}{2}$ would be to say that it is the solution of the equation $2x = 1$, that is $\frac{1}{2}$ is the quantity which when multiplied by 2 gives 1. We would then use the notation $x = \frac{p}{q}$ to indicate that the numerator p is the right hand side and the denominator q is the factor on the left hand side in the equation $qx = p$. We could equally well think of $x = \frac{p}{q}$ as a *pair*, or more precisely as an *ordered pair* $x = (p, q)$ with a *first component* p and a *second component* q representing the right hand side and the left hand side factor of the equation $qx = p$ respectively. Note that the notation $\frac{p}{q}$ is nothing but an alternative way of ordering the pair of integers p and q with an "upper" p and a "lower" q; the horizontal bar in $\frac{p}{q}$ separating p and q is just a counterpart of the comma separating p and q in (p, q).

We could now directly identify some of these pairs (p, q) or "new things" with already known objects. Namely, a pair (p, q) with $q = 1$ would be identified with the integer p since in this case the equation is $1x = p$ with solution $x = p$. We could thus write $(p, 1) = p$ corresponding to writing $\frac{p}{1} = p$, as we are used to do.

Suppose now you would like to teach your roommate how to operate with rational numbers using the rules that are familiar to us who know about rational numbers, once you have conveyed the idea that a rational number is an ordered pair (p, q) with p and $q \neq 0$ integers. We could seek inspiration from the construction of the rational number $(p, q) = \frac{p}{q}$ as that thing which solves the equation $qx = p$ with p and $q \neq 0$ integers. For example, suppose we want to figure out how to multiply the rational number $x = (p, q) = \frac{p}{q}$ with the rational number $y = (r, s) = \frac{r}{s}$. We then start from the defining equations $qx = p$ and $sy = r$. Multiplying both sides, using the fact that $xs = sx$ so that $qxsy = qsxy = qs(xy)$, we find that

$$qs(xy) = pr,$$

from which we conclude that

$$xy = (pr, qs) = \frac{pr}{qs},$$

since $z = xy$ visibly solves the equation $qsz = pr$. We thus conclude the familiar rule

$$xy = \frac{p}{q} \times \frac{r}{s} = \frac{pr}{qs} \quad \text{or} \quad (p, q) \times (r, s) = (pr, qs), \tag{7.1}$$

which says that numerators and denominators are multiplied separately.

Similarly to get a clue how to add two rational numbers $x = (p, q) = \frac{p}{q}$ and $y = (r, s) = \frac{r}{s}$, we again start from the defining equations $qx = p$ and $sy = r$. Multiplying both sides of $qx = p$ by s, and both sides of $sy = r$ by q, we find $qsx = ps$ and $qsy = qr$. From these equations and the fact that for integers $qs(x + y) = qsx + qsy$, we find that

$$qs(x + y) = ps + qr,$$

which suggests that

$$x + y = \frac{p}{q} + \frac{r}{s} = \frac{ps + qr}{qs} \quad \text{or} \quad (p, q) + (r, s) = (ps + qr, qs). \tag{7.2}$$

This gives the familiar way of adding rational numbers by using a common denominator.

We further note that for $s \neq 0$, $qx = p$ if and only if $sqx = sp$, (see Problem 5.5). Since the two equations $qx = p$ and $sqx = sp$ have the same solution x,

$$\frac{p}{q} = x = \frac{sp}{sq} \quad \text{or} \quad (p, q) = (sp, sq). \tag{7.3}$$

This says that a common nonzero factor s in the numerator and the denominator may be cancelled out or, vice versa introduced.

With inspiration from the above calculations, we may now *define* the rational numbers to be the ordered pairs (p, q) with p and $q \neq 0$ integers, and we decide to write $(p, q) = \frac{p}{q}$. Inspired by (7.3), we *define* $(p, q) = (sp, sq)$ for $s \neq 0$, thus considering (p, q) and (sp, sq) to be (two representatives of) one and the same rational number. For example, $\frac{6}{4} = \frac{3}{2}$.

We next *define* the operations of multiplication $\times$ and addition $+$ of rational numbers by (7.1) and (7.2). We may further identify the rational number $(p, 1)$ with the integer p, since p solves the equation $1x = p$. We can thus view the rational numbers as an *extension* of the integers, in the same way that the integers are an extension of the natural numbers. We note that $p+r = (p, 1)+(r, 1) = (p+r, 1) = p+r$ and $pr = (p, 1)\times(r, 1) = (pr, 1) = pr$, and thus addition and multiplication of the rational numbers that can be identified with integers is performed as before.

We can also define division $(p,q)/(r,s)$ of the rational number (p,q) by the rational number (r,s) with $r \neq 0$, as the solution x of the equation $(r,s)x = (p,q)$. Since $(r,s)(ps,qr) = (rps,sqr) = (p,q)$,

$$x = (p,q)/(r,s) = \frac{(p,q)}{(r,s)} = (ps,qr),$$

which we can also write as

$$\frac{\frac{p}{q}}{\frac{r}{s}} = \frac{ps}{qr}.$$

Finally, we may *order* the rational numbers as follows. We *define* the rational number (p,q) (with $q \neq 0$) to be positive, writing $(p,q) > 0$ whenever p and q have the same sign, and for two rational numbers (p,q) and (r,s) we write $(p,q) < (r,s)$ whenever $(r,s) - (p,q) > 0$. Note the difference can be computed as $(r,s) - (p,q) = (qr - sp, sq)$ because $-(p,q)$ is just a convenient notation for $(-p,q)$. Note also that $-(p,q) = (-p,q) = (p,-q)$, which we recognize as

$$-\frac{p}{q} = \frac{-p}{q} = \frac{p}{-q}.$$

The *absolute value* $|r|$ of a rational number $r = (p,q) = \frac{p}{q}$ is defined as for natural numbers by

$$|r| = \begin{cases} r & \text{if } r \geq 0, \\ -r & \text{if } r < 0. \end{cases} \tag{7.4}$$

where as above $-r = -(p,q) = -\frac{p}{q} = \frac{-p}{q} = \frac{p}{-q}$.

We can now verify all the familiar rules for computing with rational numbers by using the rules for integers already established.

Of course we use x^n with x rational and n a natural number to denote the product of n factors x. We also write

$$x^{-n} = \frac{1}{x^n}$$

for natural numbers n and $x \neq 0$. Defining $x^0 = 1$ for x rational, we have defined x^n for x rational n integer, with $x \neq 0$ if $n < 0$.

We finally check that we can indeed solve equations of the form $qx = p$, or $(q,1)x = (p,1)$, with $q \neq 0$ and p integers. The solution is $x = (p,q)$ since $(q,1)(p,q) = (qp,q) = (p,1)$.

So we have *constructed* the rational numbers from the integers in the sense that we view each rational number $\frac{p}{q}$ as an ordered pair (p,q) of integers p and $q \neq 0$ and we have specified how to compute with rational numbers using the rules for computing with integers.

We note that any quantity computed using addition, subtraction, multiplication, and division of rational numbers (avoiding division by zero) always produces another rational number. In the language of mathematicians, the set of rational numbers is "closed" under arithmetic operations, since these operations do not lead out of the set. Hopefully, your (receptive) roommate will now be satisfied.

7.3 On the need for rational numbers

The need of using rational numbers is made clear in early school years. The integers alone are too crude an instrument and we need fractions to reach a satisfactory precision. One motivation comes from our daily experience with measuring quantities of various sorts. When creating a set of standards for measuring quantities, such as the English foot-pound system or the metric system, we choose some arbitrary quantities to mark as the unit measurement. For example, the meter or the yard for distance, the pound or the kilogram for weight, the minute or second for time. We measure everything in reference to these units. But rarely does a quantity measure out to be an even number of units and so we are forced to deal with fractions of the units. The only possible way to avoid this would be to pick extremely small units (like the Italian lire), but this is impractical. We even give names to some particular units of fractions; centimeters are $1/100$ of meters, millimeters are $1/1000$ of a meter, inches are $1/12$ of foot, ounces are $1/16$ of a pound, and so on.

Consider the problem of adding 76 cm to 5 m. We do this by changing the meters into centimeters, 5 m = 500 cm, then adding to get 576 cm. But this is the same thing as finding a common denominator for the two distances in terms of a centimeter, i.e. $1/100$ of a meter, and adding the result.

7.4 Decimal expansions of rational numbers

The most useful way to represent a rational number is in the form of a decimal expansion, such as $1/2 = 0.5$, $5/2 = 2.5$, and $5/4 = 1.25$. In general, a *finite decimal expansion* is a number of the form

$$\pm p_m p_{m-1} \cdots p_2 p_1 p_0 . q_1 q_2 \cdots q_n, \tag{7.5}$$

where the *digits* p_m, p_{m-1}, .., p_0, q_0, .., q_n are each equal to one of the natural numbers $\{0, 1, \cdots, 9\}$ while m and n are natural numbers. The decimal expansion (7.5) is a shorthand notation for the number

$$\begin{aligned} \pm\, p_m 10^m + p_{m-1} 10^{m-1} + \cdots + p_1 10^1 + p_0 10^0 \\ + q_1 10^{-1} + \cdots + q_{n-1} 10^{-(n-1)} + q_n 10^{-n}. \end{aligned}$$

For example

$$432.576 = 4 \times 10^2 + 3 \times 10^1 + 2 \times 10^0 + 5 \times 10^{-1} + 7 \times 10^{-2} + 6 \times 10^{-3}.$$

The integer part of the decimal number (7.5) is $p_m p_{m-1} \cdots p_1 p_0$, while the decimal or fractional part is $0.q_1 q_2 \cdots q_n$. For example, $432.576 = 432 + 0.576$.

The decimal expansion is computed by continuing the long division algorithm "past" the decimal point rather than stopping when the remainder is found. We illustrate in Fig. 7.1.

```
       47.55
   ----------
40 | 1902.000
     160
     -------
      302
      280
     -------
       22.0
       20.0
       ----
        2.00
        2.00
        ----
         .00
```

FIGURE 7.1. Using long division to obtain a decimal expansion.

A finite decimal expansion is necessarily a rational number because it is a sum of rational numbers. This can also be understood by writing $p_m p_{m-1} \cdots p_1 p_0.q_1 q_2 \cdots q_n$ as the quotient of the integers $p_m p_{m-1} \cdots p_1 p_0 q_1 q_2 \cdots q_n$ and 10^n:

$$p_m p_{m-1} \cdots p_1 p_0.q_1 q_2 \cdots q_n = \frac{p_m p_{m-1} \cdots p_1 p_0.q_1 q_2 \cdots q_n}{10^n},$$

like $432.576 = 432576/10^3$.

7.5 Periodic decimal expansions of rational numbers

Computing decimal expansions of rational numbers using long division leads immediately to an interesting observation: some decimal expansions do not "stop". In other words, some decimal expansions are never-ending, that is contain an *infinite* number of nonzero decimal digits. For example, the solution to the equation $15x = 10$ in the Dinner Soup model is $x = 2/3 = .666\cdots$. Further, $10/9 = 1.11111\cdots$, as displayed in Fig. 7.2. The word "infinite" is here to indicate that the decimal expansion contin-

$$
\begin{array}{r@{\,}l}
 & 1.1111... \\
9 & \overline{)\,10.0000...} \\
 & \underline{9} \\
 & 1.0 \\
 & \underline{.9} \\
 & .10 \\
 & \underline{.09} \\
 & .010 \\
 & \underline{.009} \\
 & .0010
\end{array}
$$

FIGURE 7.2. The decimal expansion of $10/9$ never stops.

ues without ever stopping. We can find many examples of infinite decimal expansions:

$$
\begin{aligned}
\frac{1}{3} &= .3333333333\cdots \\
\frac{2}{11} &= .18181818181818\cdots \\
\frac{4}{7} &= .571428571428571428571428\cdots
\end{aligned}
$$

We conclude that the system of rational numbers $\frac{p}{q}$ with p and $q \neq 0$ integers, and the decimal system, don't fit completely. To express certain rational numbers decimally is impossible with only a finite number of decimals, unless we are prepared to accept some imprecision.

We note that in all the above examples of infinite decimal expansions, the digits in the decimal expansion begin to repeat after some point. The digits in $10/9$ and $1/3$ repeat in each entry, the digits in $2/11$ repeat after every two entries, and the digits in $4/7$ repeat after every six entries. We say that these decimal expansions are *periodic*.

In fact, if we consider the process of long division in computing the decimal expansion of p/q, then we realize that the decimal expansion of any rational number must either be finite (if the remainder eventually becomes zero), or periodic (if the remainder is never zero). To see that these are the only alternatives, we assume that the expansion is not finite. At every stage in the division process the remainder will then be nonzero, and disregarding the decimal point, the remainder will correspond to a natural number r satisfying $0 < r < q$. In other words, remainders can take at most $q - 1$ different forms. Continuing long division at most q steps must thus leave a remainder, whose digits have come up at least once before. But after that first repetition of remainder, the subsequent remainders will repeat in the same way and thus the decimal expansion will eventually be periodic.

The periodic pattern of a rational number may take a long time to begin repeating. We give an example:

$$
\begin{aligned}
\frac{1043}{439} = 2.&37585421412300683371298405466970387243735763097\\
&94988610478359908883826879271070615034168564920\\
&27334851936218678815489749430523917995444191343\\
&96355353075170842824601366742596810933940774487\\
&4715261958997722095671981776765\quad 37585421412300\\
&68337129840546697038724373576309794988610478359\\
&90888382687927107061503416856492027334851936218\\
&67881548974943052391799544419134396355353075170\\
&84282460136674259681093394077448747152619589977\\
&22095671981776765\ \cdots
\end{aligned}
$$

Once a periodic pattern of the decimal expansion of a rational number has developed, then we may consider the complete decimal expansion to be known in the sense that we can give the value of any decimal of the expansion without having to continue the long division algorithm to that decimal. For example, we are sure that the 231th digit of $10/9 = 1.111\cdots$ is 1, and the 103th digit of $.56565656\cdots$ is 5.

A rational number with an infinite decimal expansion cannot be exactly represented using a finite decimal expansion. We now seek to consider the error committed by truncating an infinite periodic expansion to a finite one. Of course, the error must be equal to the number corresponding to the decimals left out by truncating to a finite expansion. For example, truncating after 3 decimals, we would have

$$\frac{10}{9} = 1.111 + 0.0001111\cdots,$$

with the error equal to $0.0001111\cdots$, which certainly must be less than 10^{-3}. Similarly, truncating after n decimals, the error would be less than 10^{-n}.

However, since this discussion directly involves the infinite decimal expansion left out by truncation, and since we have so far not specified how to operate with infinite decimal expansions, let us approach the problem from a somewhat different angle. Denoting the decimal expansion of $10/9$ truncated after n decimals by $1.1\cdots 1_n$ (that is with n decimals equal to 1 after the point), we have

$$1.11\cdots 11_n = 1 + 10^{-1} + 10^{-2} + \cdots + 10^{-n+1} + 10^{-n}.$$

Computing the sum on the right hand side using the formula (6.3) for a geometric sum, we have

$$1.11\cdots 11_n = \frac{1-10^{-n-1}}{1-0.1} = \frac{10}{9}\,(1-10^{-n-1}), \tag{7.6}$$

and thus

$$\frac{10}{9} = 1.11\cdots 11_n + \frac{10^{-n}}{9}. \tag{7.7}$$

The error committed by truncation is thus $10^{-n}/9$, which we can bound by 10^{-n} to simplify. The error $10^{-n}/9$ will get as small as we please by taking n large enough, and thus we can make $1.11\cdots 11_n$ as close as we like to $10/9$ by taking n large enough. This leads us to interpreting

$$\frac{10}{9} = 1.11111111\cdots$$

as meaning that we can make the numbers $1.111\cdots 1_n$ as close as we like to $10/9$ by taking n large. In particular, we would have

$$|\frac{10}{9} - 1.11\cdots 11_n| \leq 10^{-n}.$$

Taking sufficiently many decimals in the never ending decimal expansion of $\frac{10}{9}$ makes the error smaller than any given positive number.

We give another example before considering the general case. Computing we find that $2/11 = .1818181818\cdots$. Taking the first m pairs of the digits 18, we get

$$\begin{aligned}
.1818\cdots 18_m &= \frac{18}{100} + \frac{18}{10000} + \frac{18}{1000000} + \cdots + \frac{18}{10^{2m}} \\
&= \frac{18}{100}\left(1 + \frac{1}{100} + \frac{1}{100^2} + \cdots + \frac{1}{100^{m-1}}\right) \\
&= \frac{18}{100}\,\frac{1-(100^{-1})^m}{1-100^{-1}} = \frac{18}{100}\,\frac{100}{99}(1-100^{-m}) \\
&= \frac{2}{11}(1-100^{-m}).
\end{aligned}$$

that is

$$\frac{2}{11} = 0.1818\cdots 18_m + \frac{2}{11}100^{-m},$$

so that

$$|\frac{2}{11} - 0.1818\cdots 18_m| \leq 100^{-m}.$$

We thus interpret $2/11 = .1818181818\cdots$ as meaning that we can make the numbers $.1818\cdots 18_m$ as close as we like to $2/11$ by taking m sufficiently large.

We now consider the general case of an infinite periodic decimal expansion of the form

$$p = .q_1q_2\cdots q_nq_1q_2\cdots q_nq_1q_2\cdots q_n\cdots,$$

where each period consists of the n digits $q_1\cdots q_n$. Truncating the decimal expansion after m periods, we get using (6.3), as

$$\begin{aligned} p_m &= \frac{q_1q_2\cdots q_n}{10^n} + \frac{q_1q_2\cdots q_n}{10^{n2}} + \cdots + \frac{q_1q_2\cdots q_n}{10^{nm}} \\ &= \frac{q_1q_2\cdots q_n}{10^n}\left(1 + \frac{1}{10^n} + \frac{1}{(10^n)^2} + \cdots + \frac{1}{(10^n)^{m-1}}\right) \\ &= \frac{q_1q_2\cdots q_n}{10^n}\frac{1-(10^{-n})^m}{1-10^{-n}} = \frac{q_1q_2\cdots q_n}{10^n-1}\left(1-(10^{-n})^m\right), \end{aligned}$$

that is

$$\frac{q_1q_2\cdots q_n}{10^n-1} = p_m + \frac{q_1q_2\cdots q_n}{10^n-1}10^{-nm},$$

so that

$$\left|\frac{q_1q_2\cdots q_n}{10^n-1} - p_m\right| \le 10^{-nm}.$$

We conclude that we may interpret

$$p = \frac{q_1q_2\cdots q_n}{10^n-1}$$

to mean that the difference between the truncated decimal expansion p_m of p and $q_1q_2\cdots q_n/(10^n-1)$ can be made smaller than any positive number by taking the number of periods m large enough, that is by taking more digits of p into account. Thus, we may view p to be equal to a rational number, namely $p = q_1q_2\cdots q_n/(10^n-1)$.

EXAMPLE 7.1. $0.123123123\cdots$ is the same as the rational number $\frac{123}{99}$, and $4.121212\cdots$ is the same as $4+\frac{12}{9} = \frac{4\times 9+12}{9} = \frac{48}{9}$.

We conclude that each infinite periodic decimal expansion may be considered to be equal to a rational number, and vice versa. We may thus summarize the discussion in this section as the following fundamental theorem.

Theorem 7.1 *The decimal expansion of a rational number is periodic. A periodic decimal expansion is equal to a rational number.*

7.6 Set notation

We have already encountered several examples of *sets*, for example the set $\{1, 2, 3, 4, 5\}$ of the first 5 natural numbers, and the (infinite) set $\mathbb{N} =$

$\{1, 2, 3, 4, \cdots\}$ of all (possible) natural numbers. A set is defined by its *elements*. For example, the set $A = \{1, 2, 3, 4, 5\}$ consists of the elements $1, 2, 3, 4$ and 5. To denote that an object is an element of a set we use the symbol $\in$, for example $4 \in A$. We further have that $7 \in \mathbb{N}$ but $7 \notin A$. To define a set we have to somehow specify its elements. In the two given examples we could accomplish this by simply listing its elements within the embracing set indicators { and }. As we encounter more complicated sets we have to somewhat develop our notation. One convenient way is to specify the elements of a set through some relevant property. For example $A = \{n \in \mathbb{N} : n \leq 5\}$, to be interpreted as "the set of natural numbers n such that $n \leq 5$". For another example, the set of *odd* natural numbers could be specified as $\{n \in \mathbb{N} : n \text{ odd}\}$ or $\{n \in \mathbb{N} : n = 2j - 1 \text{ for some } j \in \mathbb{N}\}$. The colon : is here interpreted as "such as".

Given sets A and B, we may construct several new sets. In particular, we denote by $A \cup B$ the *union* of A and B consisting of all elements which belong to at least one of the sets A and B, and by $A \cap B$ the *intersection* of A and B consisting of all elements which belong to both A and B. Further $A \backslash B$ denotes the set of elements in A which do *not* belong to B, which may be interpreted as "subtracting" B (or rather $B \cup A$) from A, see Fig. 7.3

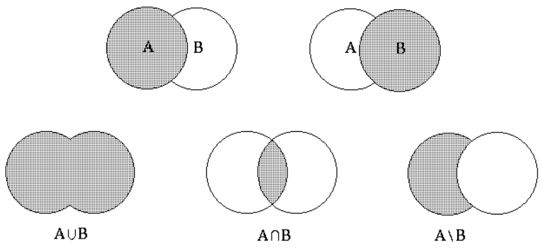


FIGURE 7.3. The sets $A \cup B$, $A \cap B$ and $A \backslash B$.

We further denote by $A \times B$ the *product set* of A and B which is the set of all possible *ordered pairs* (a, b) where $a \in A$ and $b \in B$.

EXAMPLE 7.2. If $A = \{1, 2, 3\}$ and $B = \{3, 4\}$, then $A \cup B = \{1, 2, 3, 4\}$, $A \cap B = \{3\}$, $A \backslash B = \{1, 2\}$ and $A \times B = \{(1, 3), (1, 4), (2, 3), (2, 4), (3, 3), (3, 4)\}$.

7.7 The set $\mathbb{Q}$ of all rational numbers

It is common to use $\mathbb{Q}$ to denote the set of all possible rational numbers, that is, the set of numbers x of the form $x = p/q = (p, q)$, where p and $q \neq 0$ are integers. We often omit the "possible" and just say that $\mathbb{Q}$ denotes the set of rational numbers, which we can write as

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

We can also describe $\mathbb{Q}$ as the set of finite or periodic decimal expansions.

7.8 The rational number line and intervals

Recall that we represent the integers using the integer number line, which consists of a line on which we mark regularly spaced points. We can also use a line to represent the rational numbers. We begin with the integer number line and then add the rational numbers that have one decimal place:

$$-\cdots, -1, -.9, -.8, \cdots, -.1, 0, .1, .2, \cdots, .9, 1, \cdots.$$

Then we add the rational numbers that have two decimal places:

$$-\cdots, -.99, -.98, \cdots, -.01, 0, .01, .02, \cdots, 98, .99, 1, \cdots.$$

Then onto the rational numbers with $3, 4, \cdots$ decimal places. We illustrate in Fig. 7.4. We see that there are quickly so many points to plot that the

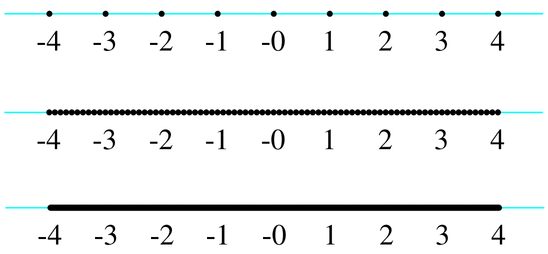


FIGURE 7.4. Filling in the rational number line between -4 and 4 starting with integers, rationals with one digit, and rationals with two digits, and so on.

number line looks completely solid. A solid line would mean that every number is rational, something we discuss later. But in any case, a drawing of a number line appears solid. We call this the *rational number line.*

For given rational numbers a and b with $a \leq b$ we say that the rational numbers x such that $a \leq x \leq b$ is a *closed interval* and we denote the interval by $[a, b]$. We also write

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

The points a and b are called the *endpoints* of the interval. Similarly we define *open* (a,b) and *half-open* intervals $[a,b)$ and $(a,b]$ by

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

$$[a,b) = \{x \in \mathbb{Q} : a \leq x < b\}, \text{ and } (a,b] = \{x \in \mathbb{Q} : a < x \leq b\}.$$

In an analogous way, we write all the rational numbers larger than a number a as

$$(a,\infty) = \{x \in \mathbb{Q} : a < x\} \text{ and } [a,\infty) = \{x \in \mathbb{Q} : a \leq x\}.$$

We write the set of numbers less than a in a similar way. We also represent intervals graphically by marking the points on the rational line segment, as we show in Fig. 7.5. Note how we use an open circle or a closed circle to mark the endpoints of open and closed intervals.

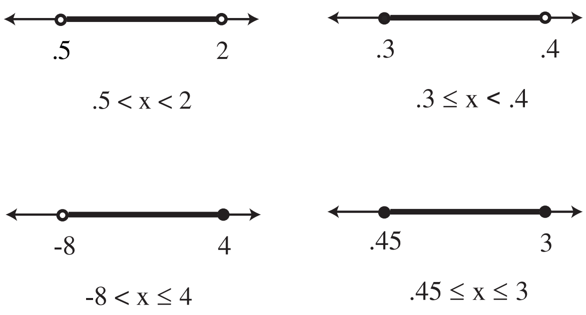


FIGURE 7.5. Various rational line intervals.

7.9 Growth of bacteria

We now present a model from biology related to population dynamics requiring the use of rational numbers.

Certain bacteria cannot produce some of the amino acids they need for the production of protein and cell reproduction. When such bacteria are cultured in growth media containing sufficient amino acids, then the population doubles in size at a regular time interval, say on the order of an hour. If P_0 is the initial population at the current time and P_n is the population after n hours, then we have

$$P_n = 2P_{n-1} \tag{7.8}$$

for $n \geq 1$. This model is similar to the model (6.5) we used to describe the insect population in Model 6.2. If the bacteria can keep growing in this way, then we know from that model that $P_n = 2^n P_0$. However if there is a limited amount of amino acid, then the bacteria begin to compete for the resource. As a result, the population will no longer double every hour. The question is what happens to the bacteria population as time increases? Does it keep increasing, does it decrease to zero (die out), or does it tend to some constant value for example?

To model this, we allow the proportionality factor 2 in (7.8) to vary with the population in such a way that it decreases as the population increases. For example, we assume there is a constant $K > 0$ such that the population at hour n satisfies

$$P_n = \frac{2}{1 + P_{n-1}/K} P_{n-1}. \tag{7.9}$$

With this choice, the proportionality factor $2/(1 + P_{n-1}/K)$ is always less than 2 and clearly decreases as P_{n-1} increases. We emphasize that there are many other functions that have this behavior. The right choice is the one that gives results that match experimental data from the laboratory. It turns out that the choice we have made does fit experimental data well and (7.9) has been used as a model not only for bacteria but also for certain human populations as well as for fisheries.

We now seek a formula expressing how P_n depends on n. We define $Q_n = 1/P_n$, then (7.9) implies (check this!) that

$$Q_n = \frac{Q_{n-1}}{2} + \frac{1}{2K}.$$

Now we use induction as we did for the insect model:

$$\begin{aligned}
Q_n &= \frac{1}{2}Q_{n-1} + \frac{1}{2K} \\
&= \frac{1}{2^2}Q_{n-2} + \frac{1}{2K} + \frac{1}{4K} \\
&= \frac{1}{2^3}Q_{n-3} + \frac{1}{2K} + \frac{1}{4K}\frac{1}{8K} \\
&\;\;\vdots \\
&= \frac{1}{2^n}Q_0 + \frac{1}{2K}\left(1 + \frac{1}{2} + \cdots + \frac{1}{2^{n-1}}\right)
\end{aligned}$$

With each hour that passes, we add another term onto the sum giving Q_n while we want to figure out what happens to Q_n as n increases. Using the formula for the sum of the geometric series (6.3), which turns out to hold for the sum of rational numbers as well as for integers, we find

$$P_n = \frac{1}{Q_n} = \frac{1}{\frac{1}{2^n}Q_0 + \frac{1}{K}\left(1 - \frac{1}{2^n}\right)}. \tag{7.10}$$

7.10 Chemical equilibrium

The solubility of ionic precipitates is an important issue in analytical chemistry. For the equilibrium

$$A_x B_y \rightleftharpoons x A^{y+} + y B^{x-} \tag{7.11}$$

for a saturated solution of slightly soluble strong electrolytes, the solubility product constant is given by

$$K_{sp} = [A^{y+}]^x [B^{x-}]^y. \tag{7.12}$$

The solubility product constant is useful for predicting whether or not a precipitate can form in a given set of conditions and the solubility of an electrolyte for example.

We will use it to determine the solubility of $Ba(IO_3)_2$ in a .020 mole/liter solution of KIO_3 :

$$Ba(IO_3)_2 \rightleftharpoons Ba^{2+} + 2\,IO_3^-$$

given that the K_{sp} for $Ba(IO_3)_2$ is 1.57×10^{-9}. We let S denote the solubility of $Ba(IO_3)_2$. By a mass law, we know that $S = [Ba^{2+}]$ while iodate ions come from both the KIO_3 and the $Ba(IO_3)_2$. The total iodate concentration is the sum of these contributions,

$$[IO_3^-] = (.02 + 2S).$$

Substituting these into (7.12), we get the equation

$$S\,(.02 + 2S)^2 = 1.57 \times 10^{-9}. \tag{7.13}$$

Chapter 7 Problems

7.1. Explain to your roommate what rational numbers are and how to manipulate them. Change roles in this game.

7.2. Prove the commutative, associative and distributive law for rational numbers.

7.3. Verify the commutative and distributive rules for addition and multiplication of rational numbers from the given definitions of addition and multiplication.

7.4. Using the usual definitions for multiplication and additions of rational numbers show that if r, s and t are rational numbers, then $r(s + t) = rs + rt$.

7.5. Determine the set of x satisfying the following inequalities:

(a) $|3x-4| \le 1$ (b) $|2-5x| < 6$

(c) $|14x-6| > 7$ (d) $|2-8x| \ge 3$

7.6. Verify that for rational numbers r, s, and t

$$|s-t| \le |s| + |t|, \tag{7.14}$$

$$|s-t| \le |s-u| + |t-u|, \tag{7.15}$$

and

$$|st| = |s|\,|t|. \tag{7.16}$$

7.7. A person running on a large ship runs 8.8 feet/second while heading toward the bow while the ship is moving at 16 miles/hour. What is the speed of the runner relative to a stationary observer? Interpret the computation giving the solution as finding a common denominator.

7.8. Compute decimal expansions for (a) 3/7, (b) 2/13, and (c) 5/17.

7.9. Compute decimal expansions for (a) 432/125 and (b) 47.8/80.

7.10. Find rational numbers corresponding to the decimal expansions (a) $42424242\cdots$, (b) $.881188118811\cdots$, and (c) $.4290542905\cdots$.

7.11. Represent the following sets as parts of the rational number line:

(a) $\{x \in \mathbb{Q} : -3 < x\}$
(b) $\{x \in \mathbb{Q} : -1 < x \le 2 \text{ and } 0 < x < 4\}$
(c) $\{x \in \mathbb{Q} : -1 \le x \le 3 \text{ or } -2 < x < 2\}$
(d) $\{x \in \mathbb{Q} : x \le 1 \text{ or } x > 2\}$.

7.12. Find an equation for the number of milligrams of $Ba(IO_3)_2$ that can be dissolved in 150 ml of water at 25° C with $K_{sp} = 1.57 \times 10^{-9}$ moles^2/liter^3. The reaction is

$$Ba(IO_3)_2 \rightleftharpoons Ba^{2+} + 2\,IO_3^-$$

7.13. You invest some money in a bond that yields 9% interest each year. Assuming that you invest any money you make from interest in more bonds for an initial investment of $\$C_0$, write down a model giving the amount of money you have after n years. View the growth of your capital with n using *MATLAB*© for example.

8
Pythagoras and Euclid

(1) At its deepest level, reality is mathematical in nature.
(2) Philosophy can be used for spiritual purification.
(3) The soul can rise to union with the divine.
(4) Certain symbols have a mystical significance.
(5) All brothers of the order should observe strict loyalty and secrecy.
(Beliefs of Pythagoras)

8.1 Introduction

In this chapter, we discuss a couple of basic useful facts from *geometry*. In doing so, we make connections to the origins of mathematics in ancient Greece 2500 years ago and in particular to two portal figures: Pythagoras and Euclid, see Fig. 8.1. In their work, we find roots of most of the topics we will meet below.

8.2 Pythagoras theorem

You have probably heard about *Pythagoras theorem* for a triangle with a *right angle* since it is a very fundamental and important result of geometry. Recall, we used Pythagoras theorem in the Muddy Yard model for example. Pythagoras theorem states that if the sides next to the right angle have lengths a and b and the side opposite the right angle (the hypotenuse) has

FIGURE 8.1. Pythagoras and Euclid at work.

length c, then (see Fig. 8.2),

$$c^2 = a^2 + b^2.$$

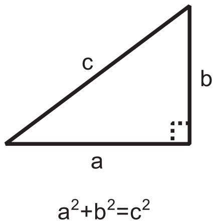


FIGURE 8.2. Pythagoras' theorem.

Suppose that our roommate has never heard about this result, as unlikely as it may be. How can we convince her/him that Pythagoras theorem is true? That is, how can we *prove* it? Well, we could argue as follows: Construct a rectangle surrounding the triangle by drawing a line parallel to the hypotenuse through the right angle corner and lines at right angles to the hypotenuse through the other two corners, see Fig. 8.3. We have now three triangles, two new adjoined to the original one. All these triangles have the same form, that is they have the same angles: one right angle of $90°$, one angle of size α and one of size β, see Fig. 8.3. This is because

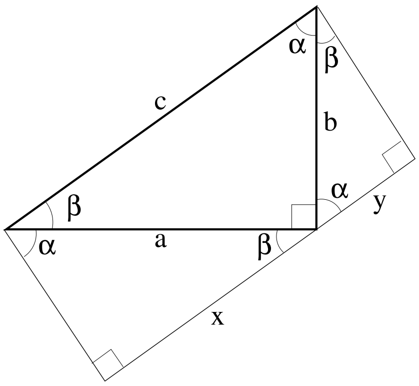


FIGURE 8.3. Proof of Pythagoras theorem.

$\alpha + \beta = 90°$, since the sum of the angles of a triangle is always 180°, and a right angle is 90°. Since the three triangles have the same angles, they have the same form, or in other words, the triangles are *similar*. This means that the ratio of corresponding sides is equal for all three triangles. Using this fact twice, we see that $x/a = a/c$ and $y/b = b/c$. We conclude that $c = x + y = a^2/c + b^2/c$ or $c^2 = a^2 + b^2$, which is the statement of Pythagoras theorem.

That should convince our roommate granted that she/he is reasonably familiar with (i) the fact that the sum of the angles of a triangle is 180°, and (ii) the concept of *similar triangles*. If not, we may have to help our roommate through the following two sections.

8.3 The sum of the angles of a triangle is 180°

Consider a triangle with corners A, B and C and angles α, β and γ according to Fig. 8.4. We recall that the angle between two lines (or line

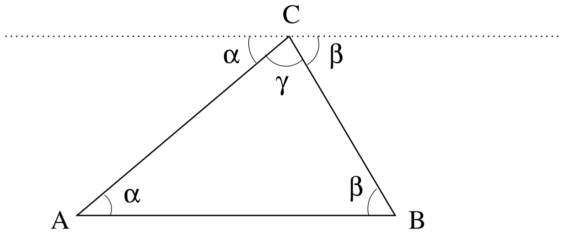


FIGURE 8.4. The angles of a triangle sum up to 180°.

segments) meeting at a point, measures how much one of the lines is rotated with respect to the other line. The most natural unit for this is "turn" or "revolution". The arm of a clock makes a full turn from 12 to 12 and a half turn from 12 to 6. When we turn the first page of our newspaper over and all the way around, we have rotated it one full turn. If we have plenty of space, like when sitting at the breakfast table alone, we just turn the page over which makes an angle of half a turn. We commonly use a system of *degrees* to measure angles, where one full turn corresponds to 360 degrees. Of course half a turn is the same as 180 degrees and quarter of a turn is 90 degrees, which is a *right* angle.

Returning to the triangle in Fig. 8.4, we would like to understand why $\alpha + \beta + \gamma = 180°$. To do this, draw a straight line parallel to the base AB through the corner C opposite to the base. The angles formed at C are α, γ and β, and their sum thus is 180°, which we wanted to show. We here use the fact that a line crossing two parallel lines crosses the two parallel lines at equal angles, see Fig. 8.5. This statement is Euclid's famous fifth *axiom*

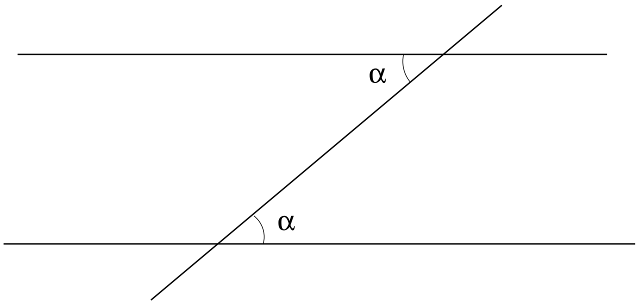


FIGURE 8.5. Two parallel lines intersected by a third one.

for geometry, which is called the *parallel axiom*. An axiom is something we take for granted without asking for any motivation. We may accept the validity of an axiom on intuitive grounds, or just accept it as a rule of a game. Euclidean geometry is based on five axioms, the first of which states that through two different points there is a unique straight line, see Fig. 8.6. In the axioms, undefined concepts like *point* and *straight line* appear. When we think of these concepts, we use our intuition from our experience of the real world. The second axiom states that a piece of a straight line can be extended to a straight line. The third axiom states that one can construct a circle with a given center and given radius and the fourth axiom states that all right angles are equal.

Euclidean geometry concerns *plane* geometry, which may be thought of as the geometry on a *flat surface*. We may think of a soccer field as being flat, but we know that very large pieces of surface of the Earth cannot be considered to be fully flat. Geometry on curved surfaces is called *non-Euclidean geometry*. Non-Euclidean geometry has some surprising features

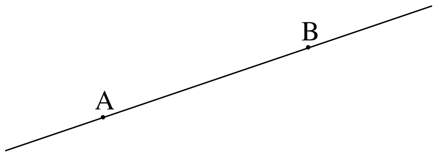


FIGURE 8.6. Euclids first axiom: through two different points there is a unique straight line.

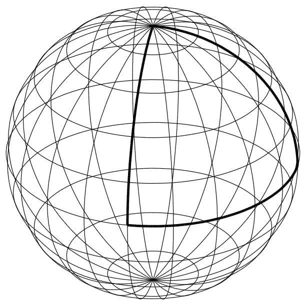
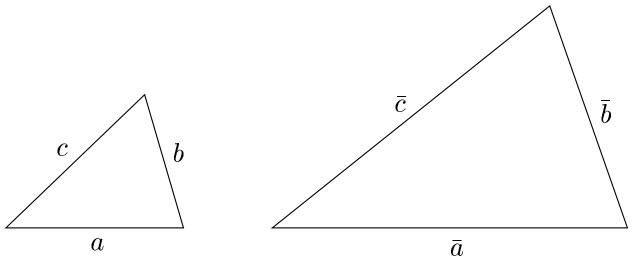


FIGURE 8.7. A "triangle" on the Earth with one corner at the North Pole and two corners at the Equator with longitude difference of 90°.

and in particular there is no parallel axiom. Consequently, the sum of the angles of a triangle on the curved surface of the Earth may be different from 180°. See Fig. 8.7, where the sum of the angles of the indicated triangle with three right angles (can you see it?) is 270°, and not at all 180°.

8.4 Similar triangles

Two triangles with the same angles are said to be *similar*. Euclid says that the ratio of corresponding sides of two similar triangles are the same. If a triangle has sides of lengths a, b and c, then a similar triangle will have side lengths $\bar{a} = ka$, $\bar{b} = kb$ and $\bar{c} = kc$, where $k > 0$ is a common *scale factor*, in other words the ratio of the corresponding side lengths is the same: $\frac{ka}{kb} = \frac{a}{b}$, $\frac{ka}{kc} = \frac{a}{c}$ and $\frac{kb}{kc} = \frac{b}{c}$. Another way of viewing this fact is to say that the angles of a triangle do not change if we change the *size* of the triangle by changing the side lengths with a common factor. This is not true in non-Euclidean geometry. If we increase the size of a (large) triangle on the surface of the Earth, then the angles will increase. We pose as a challenge to the reader the problem of proving from Euclid's axioms that similar triangles indeed have proportional sides, see Problem 8.5.



$$a/b = \bar{a}/\bar{b}$$

FIGURE 8.8. The sides of similar triangles are proportional.

8.5 When are two straight lines orthogonal?

We continue with an application of Pythagoras theorem that is of fundamental importance in both calculus and linear algebra, and which has served as the basic theoretical tool of a carpenter through the centuries. We ask the question: how can we determine if two intersecting straight lines of the Euclidean plane intersect under a right angle, that is if the two straight lines are orthogonal or not, see Fig. 8.9. This question typically comes up when constructing the foundation of a rectangular building. Assume that the two lines intersect at the point O and assume one of the

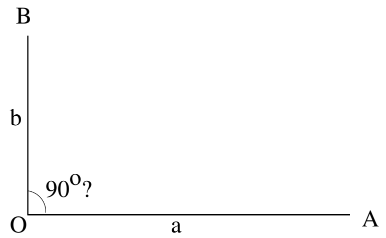


FIGURE 8.9. Are these two (pieces of) lines orthogonal?

lines passes through a point A and the other through a point B in the plane, see Fig. 8.9. We now consider the triangle AOB and ask if the angle AOB is equal to 90°, that is if the side OA is *orthogonal* to the side OB, see Fig. 8.10 We could try to check this directly with a portable right angle, but that would always leave some room for incorrect decision if the dimension of the portable right angle is much smaller than the triangle AOB itself. This would normally be the case when seeking to determine where to put the corners of a foundation of a building. Another possibility is to use Pythagoras theorem. We would expect that if

$$a^2 + b^2 = c^2 \tag{8.1}$$

where a is the length of the side OA, b is the length of side OB and c that of the side AB, see Fig. 8.10, then the angle BOA would be 90°. We

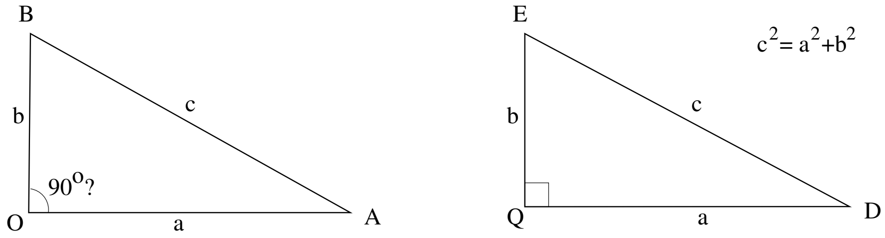


FIGURE 8.10. Test of rightangledness.

shall prove that this is so shortly, but let's first see how a carpenter would use this result in practice. He would then cut three pieces of a string, of length 3, 4 and 5 units. The beauty of these numbers is that they fit into the equation (8.1):

$$3^2 + 4^2 = 5^2. \tag{8.2}$$

If our conjecture is correct, then a triangle with sides 3, 4 and 5 will have a right angle. Assuming now that $a = 3$ and $b = 4$, we could check if the triangle AOB has a right angle by putting the string of length 5 along AB and check if $c = 5$. If so, then the angle AOB would be 90°. Using three pieces of strings of length 3, 4 and 5 units, the carpenter can thus construct a right angle. The choice of units is important in practical applications of this idea. Very short strings would be cheap but the precision would suffer, and very long strings would be impractical to handle.

Let's now go back to the conjecture that if $a^2 + b^2 = c^2$ then the triangle AOB would have a right angle. To convince ourselves we could following Euclid argue as follows: Construct a right-angled triangle DQE with the right angle at Q and with the length of DQ equal to a and the length of EQ equal to b. Then by Pythagoras theorem the length of DE squared would be equal to $a^2 + b^2$. But we assume that $c^2 = a^2 + b^2$ and thus the length of DE is c. This means that the triangles DQE and AOB would have the same side lengths and thus would be similar. Hence, the angle AOB would be equal to the angle DQE, which is equal to 90°, and we have proved that AOB is right angled.

We can use the equality $a^2 + b^2 = c^2$ as a test of orthogonality of the sides OA and OB using the trick of the carpenter, but this could still leave some uncertainty, for instance if we had to use strings of dimension very much smaller than the foundation.

Suppose now that we know the coordinates (a_1, a_2) of the corner A and the coordinates (b_1, b_2) of corner B. This could be the case if the triangle is in fact defined by specifying these coordinates, which could happen if we use a map to identify the triangle. Then we would have

$$a^2 = a_1^2 + a_2^2, \quad b^2 = b_1^2 + b_2^2, \quad c^2 = (b_1 - a_1)^2 + (b_2 - a_2)^2$$

where the last equality follows from Fig. 8.11. We conclude that

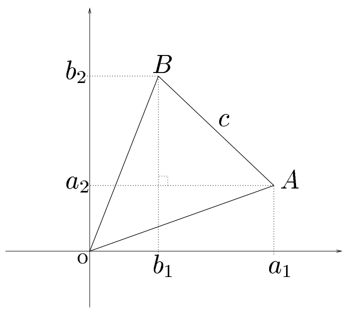


FIGURE 8.11.

$$c^2 = b_1^2 + a_1^2 - 2b_1a_1 + b_2^2 + a_2^2 - 2b_2a_2 = a^2 + b^2 - 2(a_1b_1 + a_2b_2).$$

We see that $a^2 + b^2 = c^2$ if and only if

$$a_1b_1 + a_2b_2 = 0. \tag{8.3}$$

Thus, our test of orthogonality is reduced to checking the algebraic relation $a_1b_1 + a_2b_2 = 0$. If we know the coordinates (a_1, a_2) and (b_1, b_2), this condition can be checked by multiplying numbers and is therefore not subject to taste or individual decision (up to round-off)

The magic formula (8.3) for checking orthogonality will play an important role below. It translates a geometric fact (orthogonality) into an arithmetic equality.

8.6 The GPS Navigator

GPS (Global Positioning System) is a wonderful new invention. It was developed by the U.S. Military in the 1980s, but now has important civil use. You can buy at GPS receiver for less than 200 dollars and carry it in your pocket on your hiking tour in the mountains, in your sailing boat out on the ocean, or in your car driving through the desert or the city of Los Angeles. At a press of a button on the receiver, it gives you your present coordinates: latitude and longitude as an ordered pair of numbers $(57.25, 12.60)$, where the first number (57.25) is the latitude, and the second number is the longitude (12.60). The precision may be about 10 meters. More advanced use of GPS may give you a precision as good as 1 millimeter.

Knowing your coordinates, you can identify your present position on a map and decide in what direction to proceed to come to your goal. GPS thus solves the main problem of navigation: to figure out your coordinates on a map. Before the GPS this could be very difficult and the results could be very unreliable. Determining the latitude was relatively easy if you could see the sun and measure its height over the horizon at noon using a sextant. Determining the longitude was much more difficult and would require a good clock. The main motivation for all the work that went into developing accurate clocks or chronometers in the 18th century, was to determine your longitude at sea. The results before the chronometer could be grossly inaccurate: you could believe, like Columbus, that you had come to India, while in fact you were close to America!

FIGURE 8.12. A popular GPS Navigator (Garmin 12, $200)

GPS is based on a simple (and clever) mathematical principle. We present the basic idea supposing that our world is a Euclidean plane equipped with a coordinate system. Our problem is to find our coordinates. Suppose we can measure our distance to two points A and B with known positions, which we denote by d_A and d_B. With the available information, we can say that we must be located at one of the points of intersection of the circle with center A and radius d_A and the circle with center B and radius d_B, according to Fig. 8.13. Note that Euclid's third axiom tells us that the two

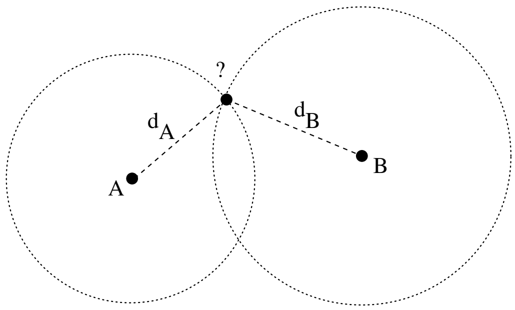


FIGURE 8.13. Positioning using GPS.

circles do exist.

To find the coordinates (x_1, x_2) of the intersection points, we have to solve the following system of two equations in the two unknowns x_1 and x_2:

$$(x_1 - a_1)^2 + (x_2 - a_2)^2 = d_A^2 \quad (x_1 - b_1)^2 + (x_2 - b_2)^2 = d_B^2.$$

If we can solve this system of equations, we can determine where we are, granted we have some extra information which tells us which of the two possible solutions applies. In three dimensions, we would have to determine our distance to three points in space with known locations and solve a corresponding system of three equations with three unknowns, corresponding to the intersection of three spheres.

GPS uses a system of 24 satellites deployed in groups of 4 in each of 6 orbital planes. Each satellite has a circular orbit of radius about 26.000 km and the orbital period is 12 hours. Each satellite has an accurate clock and there is a surveying system that keeps track of the positions of the satellites at each time instant. A GPS receiver receives a signal from each visible satellite which encodes the position and clock time of the satellite. The receiver measures the time delay, by comparing the received time with its own time, and then computes the distance to the satellite knowing the speed of light. Knowing the distance to three satellites and their positions, the GPS then computes its position as one of the two intersection points of three spheres. One of these points will be way out in space and can be eliminated if we are sure that we are on Earth. The result is our location in space specified with latitude, longitude and height above the sea. In practice, a 4th satellite is needed to calibrate the clocks of the satellites and the receiver, which is necessary to compute the distances correctly. If more than 4 satellites are visible, the GPS computes a *least squares* approximate solution to the resulting overdetermined system of equations, and the precision increases with the number of satellites.

GPS gives a coupling of geometry to arithmetics or algebra. Each physical point is space is connected to a triple of numbers representing latitude, longitude and height.

8.7 Geometric definition of sin(v) and cos(v)

Consider a right-angled triangle with an angle v, and sides of length a, b and c, as in Fig. 8.14. We recall the definitions

$$\cos(v) = \frac{a}{c} \quad \sin(v) = \frac{b}{c}.$$

Note that the values of cos(v) and sin(v) do not depend of the particular

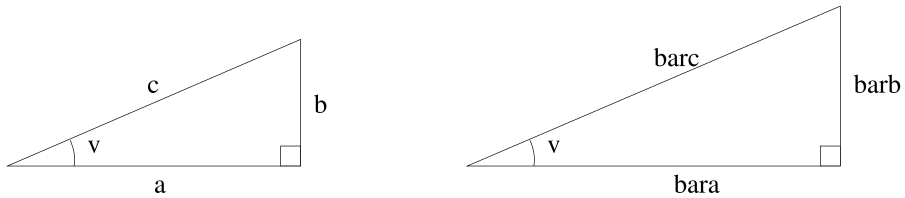


FIGURE 8.14. $\cos(v) = a/c$ and $\sin(v) = b/c$.

triangle we have used, only on v. This is because another right-angled triangle with the same angle v and sides $\bar{a}$, $\bar{b}$ and $\bar{c}$ as to the right in Fig. 8.14 would be *similar* to the one considered first, and ratios of corresponding pairs of sides in similar triangles are the same, as we have concluded before. In other words, $a/c = \bar{a}/\bar{c}$ and $b/c = \bar{b}/\bar{c}$. For example, we may use a triangle with $c = 1$ in which case $\cos(v) = a$ and $\sin(v) = b$. If we imbed such a triangle in the *unit circle* $x_1^2 + x_2^2 = 1$ as illustrated in Fig. 8.15, then $a = x_1$ and $b = x_2$. That is, $\cos(v) = x_1$ and $\sin(v) = x_2$. This imbedding also makes it possible to extend the definition of cos(v) and sin(v), and in particular, to angles v with $90^o \leq v < 180^o$.

8.8 Geometric proof of addition formulas for cos(v)

As an application of the definition of sin(v) and cos(v), we give a geometric proof of the following infamous formula,

$$\cos(\beta - \alpha) = \cos(\beta)\cos(\alpha) + \sin(\beta)\sin(\alpha) \tag{8.4}$$

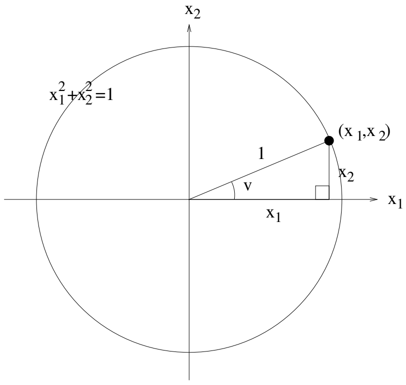


FIGURE 8.15. $\cos(v) = x_1$ and $\sin(v) = x_2$.

The proof is given in the following figure and is based on using the definitions of $\cos(v)$ and $\sin(v)$ with $v = \alpha$, $v = \beta$ and $v = \beta - \alpha$, respectively. Consider the two right-angled triangles to the left in Fig. 8.16 with a common side of length 1, and with the lengths of the other sides expressed in terms of sines and cosines of the present angles. To the right in Fig. 8.16, we have formed two other right-angled triangles in the same figure, both with an angle β. From the definition of $\cos(\beta)$ and $\sin(\beta)$, respectively, we find that the base in the larger left triangle is $\cos(\beta)\cos(\alpha)$ and the base in the smaller upper right triangle is $\sin(\beta)\sin(\alpha)$. We thus conclude that $\cos(\beta - \alpha) = \cos(\beta)\cos(\alpha) + \sin(\beta)\sin(\alpha)$.

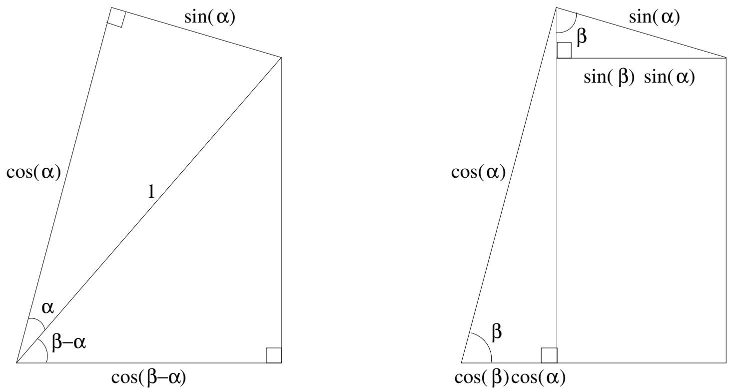


FIGURE 8.16. Why $\cos(\beta - \alpha) = \cos(\beta)\cos(\alpha) + \sin(\beta)\sin(\alpha)$.

From Fig. 8.16, we can similarly conclude that

$$\sin(\beta - \alpha) = \sin(\beta)\cos(\alpha) - \cos(\beta)\sin(\alpha). \qquad (8.5)$$

Below ,we shall redefine $\cos(v)$ and $\sin(v)$ as the solutions of certain fundamental differential equations. In particular, this will define $\cos(v)$ and $\sin(v)$ for arbitrary "angles" v, including vs greater than 180° and less than 0°, the latter corresponding to angles obtained by moving the point (x_1, x_2) clockwise from $(1, 0)$ on the unit circle in Fig. 8.15, with $\sin(v) = -\sin(-v)$ and $\cos(v) = \cos(-v)$ in accordance with the formulas $\cos(v) = x_1$ and $\sin(v) = x_2$ above.

8.9 Remembering some area formulas

We remind the reader about the following well-known formulas for the areas of rectangles, triangles, and circles, shown in Fig. 8.17. Below we will come

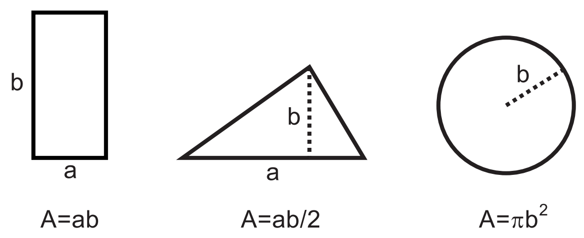


FIGURE 8.17. Some common formulas for areas A.

back to these basic formulas and motivate them more closely.

8.10 Greek mathematics

Greek mathematics is dominated by the schools of Pythagoras and Euclid. Pythagoras' school is based on *numbers*, that is arithmetic, and Euclid's on *geometry*. 2000 years later in the 17th century, Descartes fused the two approaches together by creating *analytic geometry*.

Mathematics was traditionally used in Babylonia and Greece for practical calculations in astronomy and navigation et cetera, while the Pythagoreans "transformed mathematics into an education for aristocrats (free men) from being a skill of slaves". As a result experimental science and mechanics became poorly developed in the classical Greek period. Instead the principle of logical deduction was dominating.

Pythagoras was born 585 B.C. on the island Samos off the coast of Asia Minor and founded his own school on Croton, a Greek settlement in southern Italy. The Pythagorean school was a brotherhood that kept its deepest knowledge secret. The Pythagoreans associated with aristocrats and Pythagoras himself was murdered for political reasons about 497 B.C., after which his followers spread over Greece.

The Platonic school, called the Academy, was founded in Athens 387 B.C. by Plato (427-347 B.C), a great idealistic philosopher. Plato was a follower of the Pythagoreans and said that "arithmetic has a very great and elevating effect, compelling the soul to reason about abstract number, and rebelling against the introduction of visible or tangible objects into the argument..." Plato liked mathematics because it gave evidence of existence of that wonderful world of perfect objects like numbers, triangles, points, et cetera, with which Plato was so enthralled. Plate considered the real world to be imperfect and corrupt.

The discovery (we'll come back to this below) that the length of the diagonal of a square of side one, represented by the number $\sqrt{2}$, could not be expressed as the quotient of two natural numbers, that is that $\sqrt{2}$ is not a *rational number*, was a strong blow to Pythagorean's belief that the universe could be understood through relations between natural numbers. Ultimately, this led to the development of the Euclidean school based on geometry instead of arithmetic, where the irrational nature of $\sqrt{2}$ could be handled, or more precisely avoided. For Euclid, the diagonal of a square was just a geometric entity, which simply had a certain length. One did not have to express it using numbers. The length was what it was, namely the length of the diagonal of a square of side one. Similarly, Euclid "geometrized" arithmetic. For example, the product ab of two numbers a and b can be viewed as the area of the rectangle with sides a and b.

The prime example of a system based on logical deduction is the *Elements*, which is the monumental treatise on geometry in thirteen books by Euclid. The development proceeds from *axioms* and *definitions* through *theorems* derived by using the rules of logic.

A particular language has developed for expressing logical dependencies. For example, $A \Rightarrow B$ means "if A then B", that is, B follows from A, so that if A is valid then B is also valid, also expressed as "A *implies* B". Similarly $A \Leftarrow B$ means that A follows from B. If A implies B and also B implies A, we say that A and B are equivalent, expressed as $A \Leftrightarrow B$.

EXAMPLE 8.1. $x > 0 \Rightarrow x + 1 > 0$.

Book I begins with definitions and axioms, and continues with theorems on congruence, parallel lines, and proves the Pythagorean theorem. Book I-IV all treat rectilinear figures made up by straight line segments. Book V concerns proportions (!) and is considered as maybe the greatest achievements of Euclidean geometry. Book VI treats similar figures (!!). Books VII-IX concern natural numbers, and Book X goes into irrational numbers like $\sqrt{2}$. Books XI, XII and XIII concern geometry in three dimensions and the method of exhaustion connected to computing the area of curvilinear figures or volumes.

8.11 The Euclidean plane $\mathbb{Q}^2$

If we accept the parallel axiom, which in particular leads to the conclusion that the sum of the angles of a triangle is 180° as we saw above, this means that effectively we are assuming that we are working on a flat surface or a *Euclidean plane.*

We thus assume the reader has an intuitive idea of a *Euclidean plane*, to be thought of as a very large flat surface extending in all directions without any borders. For instance, like an endless flat parking place without a single car. In this Euclidean plane, which we imagine consists of *points*, we imagine a *coordinate system* consisting of two straight lines in the plane meeting at a right angle at a point which we call the *origin.* We imagine each coordinate axis to be a copy of the line of rational numbers $\mathbb{Q}$, see Fig. 8.18.

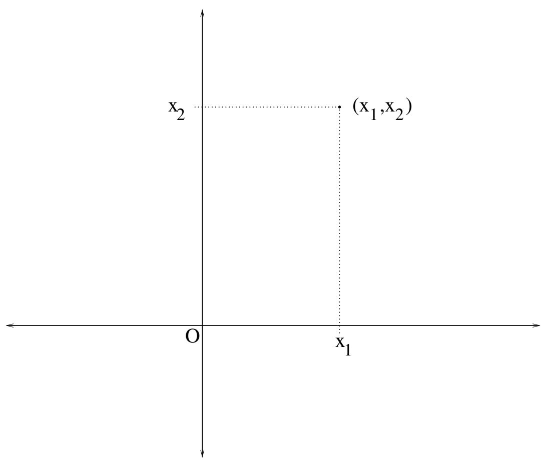


FIGURE 8.18. Coordinate system of the Euclidean plane.

We identify one coordinate axis as axis 1 and the other as coordinate axis 2. We can then associate to each *point* in the plane two *coordinates* x_1 and x_2, where x_1 represents the intersection with axis 1 of the line through the point parallel to axis 2 and vice versa. We write the coordinates as (x_1, x_2), which we think of as an *ordered pair* of rational numbers with a *first component* $x_1 \in \mathbb{Q}$ and a *second component* $x_2 \in \mathbb{Q}$. We thus seek to represent the Euclidean plane as the set $\mathbb{Q}^2 = \{(x_1, x_2) : x_1, x_2 \in \mathbb{Q}\}$, that is, the set of ordered pairs (x_1, x_2) with x_1 and x_2 rational numbers. To each point in the Euclidean plane corresponds its coordinates, which is an ordered pair of rational numbers, and to each ordered pair of rational numbers corresponds the point in the plane with those coordinates.

8.12 From Pythagoras to Euclid to Descartes

The idea is thus to associate to each geometrical point of the Euclidean plane, its coordinates as an ordered pair of rational numbers. This couples geometrical points to arithmetic, or *algebra* based on numbers, and is a most basic and fruitful connection to make. This was the idea of Descartes who followed the idea of Pythagoras to base geometry on numbers. Euclid turned this around and based numbers on geometry.

We will come back to the coupling of geometry and algebra many times below. It is probably the most basic idea of the whole of mathematics. The power comes from associating numbers to points in space, and then working with numbers instead of points. This is the "arithmetrization" of geometry of Descartes in the 17th century, which preceded Calculus and changed the history of mankind.

However, following the idea of Pythagoras and Descartes to base geometry on numbers, we will run into a quite serious difficulty which will force us to extend the rational numbers to *real numbers* including also so called *irrational numbers.* This exciting story will be unraveled below.

8.13 Non-Euclidean geometry

But, not all geometry that is useful is based on Euclid's axioms. In this century, for example, physicists have used non-Euclidean geometry to explain how the universe behaves. One of the first people to consider non-Euclidean geometry was the great mathematician Gauss.

Gauss' interest in non-Euclidean geometry gives a good picture of how his mind worked. By the time Gauss was sixteen, he had begun to seriously question Euclidean geometry. At the time that Gauss lived, Euclidean geometry had obtained an almost holy status and was held by many mathematicians and philosophers to be one of the higher truths that could never be questioned. Yet, Gauss was bothered by the fact that Euclidean geometry rested on postulates that apparently could not be proved, such as two parallel lines cannot meet. He went on to develop a theory of *non*-Euclidean geometry in which parallel lines *can* meet and this theory seemed to be as good as Euclidean geometry for describing the world. Gauss did not publish his theory, fearing too much controversy, but he decided that it should be tested. In Euclidean geometry, the sum of the angles in a triangle add up to $180°$ while in the non-Euclidean geometry this is not true. So centuries before the age of modern physics, Gauss conducted an experiment to see if the universe is "curved" by measuring the angles in the triangle made up by three mountain peaks. Unfortunately, the accuracy of his instruments was not good enough to settle the question.

Chapter 8 Problems

8.1. Which point has the coordinates $(57.25, 12.60)$?

8.2. Determine your own coordinates.

8.3. Derive (8.5) from Fig. 8.16.

8.4. Give another proof of Pythagoras' Theorem.

8.5. Prove from Euclid's axioms that the sides of two triangles with the same angles are proportional (Hint: use the 5th axiom a lot).

8.6. Prove that the three lines bisecting each angle of a triangle, intersect at a common point inside the triangle. (Hint: Decompose the given triangle into three triangles joining the corners with the point intersection of two of the bisectors).

8.7. Prove that the three lines joining each corner of a given triangle with the mid-point of the opposite side, intersect at a common point inside the triangle.

FIGURE 8.19. Classical tools of navigation: compass and sextant.

9
What is a Function?

He who loves practice without theory is like the sailor who boards ship without rudder and compass and never knows where he may cast. (Leonardo da Vinci)

All Bibles or sacred codes have been the causes of the following Errors:
1. That Man has two real existing principles, Viz: a Body & a Soul.
2. That Energy, call'd Evil, is alone from the Body; & that Reason, call'd Good, is alone from the Soul.
3. That God will torment Man in Eternity for following his Energies.
But the following Contraries to these are True:
1. Man has no Body distinct from his Soul; for that call'd Body is a portion of Soul discern'd by the five Senses, the chief inlets of Soul in this age.
2. Energy is the only life and is from the Body: and Reason is the bound or outward circumference of Energy.
3. Energy is Eternal Delight. (William Blake 1757-1827)

9.1 Introduction

The concept of a *function* is fundamental in mathematics. We already met this concept in the context of the Dinner Soup model, where the total cost was $15x$ (dollars) if the amount of beef was x (pounds). For every amount of beef x, there is a corresponding total cost $15x$. We say that the total cost $15x$ is a function of, or depends on, the amount of beef x.

The term function and the mathematical notation we use today was introduced by Leibniz (1646-1716), who said that $f(x)$, which reads "f of x", is a *function* of x if for each value of x in some prescribed set of values over which x can vary, there is assigned a unique value $f(x)$. In the Dinner Soup model $f(x) = 15x$. It is helpful to think of x as the *input*, while $f(x)$ is the corresponding *output*, so that as the value of x varies, the value of $f(x)$ varies according to the assignment. Correspondingly, we often write $x \to f(x)$ to signify that x is mapped onto $f(x)$. We also think of the function f as a "machine" that transforms x into $f(x)$:

$$x \stackrel{f}{\to} f(x),$$

see also Fig. 9.1.

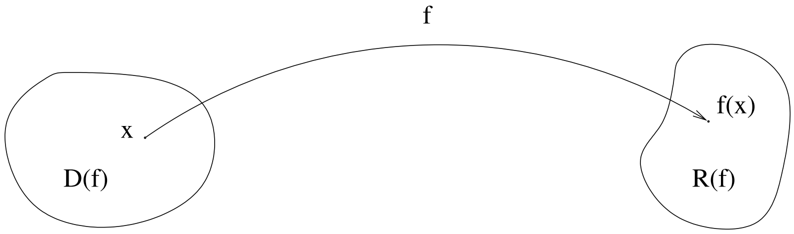


FIGURE 9.1. Illustration of $f : D(f) \to R(f)$.

We refer to x as a *variable* since x can take different values, and x is also called the *argument* of the function. The prescribed set of values over which x can vary is called the *domain* of the function f and is denoted by $D(f)$. The set of values $f(x)$ corresponding to the values of x in the domain $D(f)$, is called the *range* $R(f)$ of $f(x)$. As x varies over the domain $D(f)$, the corresponding function value $f(x)$ varies over $R(f)$. We often write this symbolically as $f : D(f) \to R(f)$ indicating that for each $x \in D(f)$ there is a value $f(x) \in R(f)$ assigned.

In the context of the Dinner Soup model with $f(x) = 15x$, we may choose $D(f) = [0, 1]$, if we decide that the amount of beef x can vary in the interval $[0, 1]$, in which case $R(f) = [0, 15]$. For each amount x of beef in the interval $[0, 1]$, there is a corresponding total cost $f(x) = 15x$ in the interval $[0, 15]$. Again: the total cost $15x$ is a function of the amount of beef x. We may also choose the domain $D(f)$ to be some other set of possible values of the amount of beef x such as $D(f) = [a, b]$, where a and b are positive rational numbers, with the corresponding range $R(f) = [15a, 15b]$, or $D(f) = \mathbb{Q}^+$ with the corresponding range $R(f) = \mathbb{Q}^+$, where $\mathbb{Q}^+$ is the set of positive rational numbers. We may even consider the function $x \to f(x) = 15x$ with $D(f) = \mathbb{Q}$ and the corresponding range $R(f) = \mathbb{Q}$, which would lead outside the Dinner Soup model since there x is non-negative. For a given assignment $x \to f(x)$, that is, a given function $f(x)$, we may thus associate different domains $D(f)$ and corresponding ranges $R(f)$ depending on the setting.

It is common to assign a variable name to the output of a function, for example we may write $y = f(x)$. Thus, the value of the variable y is given by the value $f(x)$ assigned to the variable x. We therefore call x the *independent variable* and y the *dependent variable*. The independent variable x takes on values in the domain $D(f)$, while the dependent variable y takes on values in the range $R(f)$.

Note that the names we use for the independent variable and the dependent variable for a given function f can be changed. The names x and y are common, but there is nothing special about these letters. For example, $z = f(u)$ denotes the same function if we do not change f, i.e. the function $y = 15x$ can just as well be written $z = 15u$. In both cases, to a given number x or u the function f assigns that number multiplied by 15, that is $15x$ or $15u$. Thus we refer to "the function $f(x)$" while in fact it would be more correct to just say "the function f", because f is the "name" of the function, while $f(x)$ is more like a description or definition of the function. Nevertheless we will often use the somewhat sloppy language "the function $f(x)$" because it identifies both the name of the function and its definition/description.

Example 9.1. The function $x \to f(x) = x^2$, or in short the function $f(x) = x^2$, may be considered with domain $D(f) = \mathbb{Q}^+$ and range $R(f) = \mathbb{Q}^+$, but also with domain $D(f) = \mathbb{Q}$ and again $R(f) = \mathbb{Q}^+$, or with $D(f) = \mathbb{Z}$, and $R(f) = \{0, \pm 1, \pm 2, \pm 4, ...\}$. We illustrate in Fig. 9.2.

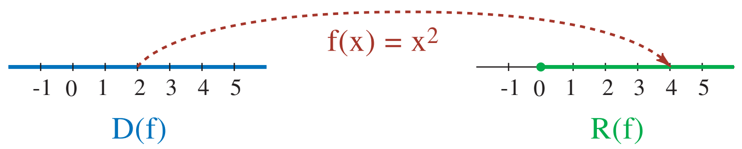


FIGURE 9.2. Illustration of $f : \mathbb{Q} \to \mathbb{Q}^+$ with $f(x) = x^2$.

Example 9.2. For the function $f(z) = z + 3$ we may choose, for example, $D(f) = \mathbb{N}$ and $R(f) = \{4, 5, 6, ...\}$, or $D(f) = \mathbb{Z}$ and $R(f) = \mathbb{Z}$.

Example 9.3. We may consider the function $f(n) = 2^{-n}$ with $D(f) = \mathbb{N}$ and $R(f) = \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, ...\}$.

EXAMPLE 9.4. For the function $x \to f(x) = 1/x$ we may choose $D(f) = \mathbb{Q}^+$ and $R(f) = \mathbb{Q}^+$. For any given x in $\mathbb{Q}^+$, the value $f(x) = 1/x$ is in $\mathbb{Q}^+$, and thus $R(f)$ is a subset of $\mathbb{Q}^+$. Correspondingly, for any given y in $\mathbb{Q}^+$ there is an x in $\mathbb{Q}^+$ with $y = 1/x$, and thus $R(f) = \mathbb{Q}^+$, that is $R(f)$ fills up the whole of $\mathbb{Q}^+$.

While the domain $D(f)$ of a function $f(x)$ often is given by the context or the nature of $f(x)$, it is often difficult to exactly determine the corresponding range $R(f)$. We therefore often interpret $f : D(f) \to B$ to mean that for each x in $D(f)$ there is an assigned value $f(x)$ that belongs to the set B. The range $R(f)$ is thus included in B, but the set B may be bigger than $R(f)$. This relieves us from figuring out exactly what set $R(f)$ is, which would be required to give $f : D(f) \to R(f)$ substance. We say that f maps $D(f)$ *onto* $R(f)$ since every element of the set $R(f)$ is of the form $f(x)$ for some $x \in D(f)$, and writing $f : D(f) \to B$ we say that f maps $D(f)$ *into* the set B.

The notation $f : D(f) \to B$ then rather serves the purpose of describing the nature or type of the function values $f(x)$, than more precisely what function values are assumed as x varies over $D(f)$ For example, writing $f : D(f) \to \mathbb{N}$ indicates that the function values $f(x)$ are natural numbers. Below we will meet functions $x \to f(x)$, where the variable x does not represent just a single number, but something more general like a pair of numbers, and likewise $f(x)$ may be a pair of numbers. Writing $f : D(f) \to B$ with the proper sets $D(f)$ and B, may contain the information that x is a number and $f(x)$ is pair of numbers. We will meet many concrete examples below.

EXAMPLE 9.5. The function $f(x) = x^2$ satisfies $f : \mathbb{Q} \to [0, \infty)$ with $D(f) = \mathbb{Q}$ and $R(f) = [0, \infty)$, but we can also write $f : \mathbb{Q} \to \mathbb{Q}$, indicating that x^2 is a rational number if x is, see Fig. 9.2.

EXAMPLE 9.6. The function

$$f(x) = \frac{x^3 - 4x^2 + 1}{(x-4)(x-2)(x+3)}$$

is defined for all rational numbers $x \neq 4, 2, -3$, so it is natural to define $D(f) = \{x \in \mathbb{Q}, x \neq 4, x \neq 2, x \neq -3\}$. It is often the case that we take the domain to be the largest set of numbers for which a function is defined. The range is hard to compute, but certainly we have $f : D(f) \to \mathbb{Q}$.

9.2 Functions in daily life

In daily life, we stumble over functions right and left. A car dealer assigns a price $f(x)$, which is a number, to each car x in his lot. Here $D(f)$ may be a set of numbers if each car is identified by a number, or $D(f)$ may be some other listing of the cars such as {Chevy85blue, Olds93pink,...}, and the range $R(f)$ is the set of all different prices of the cars in $D(f)$. When the government makes out our tax bill, it is assigning one number $f(x)$, representing the amount we owe, to another number x, representing our salary. Both the domain $D(f)$ and the range $R(f)$ in this example change a lot depending on the political winds.

Any quantity which varies over time may be viewed as a function of time. The daily maximum temperature in degrees Celsius in Stockholm during 1999 is a certain function $f(x)$ of the day x of the year, with $D(f) = \{1, 2, ..., 365\}$ and $R(f)$ normally a subset of $[-30, 30]$. The price $f(x)$ of a stock during one day of trade at the Stockholm Stock Exchange is a function of the time x of the day, with $D(f) = [10.00, 17.00]$ and $R(f)$ the range of variation of the stock price during the day. The length of women's skirts varies over the years around the level of the knee, and is supposed to be a good indicator of the variation of the economical climate. The length of a human being varies over the life time, and the thickness of the ozone layer over years.

We may also simultaneously consider several quantities depending on time, like for example the temperature $t(x)$ in degrees Celsius and wind velocity $w(x)$ in meter per second in Chicago as functions of time x, where x ranges over the month of January, and we may combine the two values $t(x)$ and $w(x)$ into a pair of numbers "$t(x)$ and $w(x)$", which we may write as $f(x) =$ "$t(x)$ and $w(x)$" or in short-hand $f(x) = (t(x), w(x))$ with the parenthesis enclosing the pair. For example writing, $f(10) = (-30, 20)$, would give the information that the 10th of January was a tough day with temperature $-30°C$ and wind 20 meter per second. From this information we could compute the adjusted temperature -50$°C$ that day taking the wind factor into account.

Likewise the input variable x could represent a pair of numbers, like a temperature and a wind speed and the output could be the adjusted temperature with the wind factor taken into account (Find the formula!).

We conclude that the input x of a function $f(x)$ may be of many different types, single numbers, pairs of numbers, triples of numbers, et cetera, as well as the output $f(x)$.

EXAMPLE 9.7. A book may consist of a set of pages numbered from 1 to N. We may introduce the function $f(n)$ defined on $D(f) = \{1, 2, \cdots, N\}$, with $f(n)$ representing the physical page with number n. In this case the range $R(f)$ is the collection of pages of the book.

EXAMPLE 9.8. A movie consists of a sequence of pictures that are displayed at the rate of 16 pictures per second. We usually watch a movie from the first to the last picture. Afterwards we might talk about different scenes in the movie, which corresponds to subsets of the totality of pictures. A very few people, like the film editor and director, might consider the movie on the level of the individual elements in the domain, that is the pictures on the film. When editing the movie, they number the picture frames $1, 2, 3, \cdots, N$ where $1, 2, ...16$ are the numbers of the pictures displayed sequentially during the first second, and N is the number of the last picture. We may then consider the movie as a function $f(n)$ with $D(f) = \{1, 2, \cdots, N\}$, which to each number n in $D(f)$ associates the picture frame with number n.

EXAMPLE 9.9. A telephone directory of the people living in a city like Göteborg is simply a printed version of the function $f(x)$ that to each person x in Göteborg with a listed number, assigns a telephone number. For example, if $x =$ Anders Andersson then $f(x) = 4631123456$ which is the telephone number of Anders Andersson. If we have to find a telephone listing, our thought is first to get the telephone book, that is the printed representation of the entire domain and range of the function f, and then to determine the image, i.e. telephone number, of an individual in the domain. In this example, we arrange the domain of individual names of people living in Göteborg in such a way that it is easy to search for a particular input. That is we list the individuals alphabetically. We could use another arrangement, say by listing individuals in order of their social security numbers.

EXAMPLE 9.10. The 1890 census (population count) in the US was performed using Herman Hollerith's (1860-1829) punched card system, where the data for each person (sex, age, address, et cet) was entered in the form of holes in certain positions on a dollarbill size card, which could then be read automatically by a machine using a system of pins connecting electrical circuits through the holes, see Fig. 9.3. The total population was found to be 62.622.250 after a processing time of three months with the Hollerith system instead of the projected 2 years. Evidently, we may view the Hollerith system as a function from the set of all 1890 US citizens to the deck of punched cards. To further exploit his system Hollerith founded the Tabulating Machine Company, which was renamed International Business Machines Corporation IBM in 1924.

.

There is one important aspect of all the three above examples, book, movie and directory, not captured viewing these objects as certain functions $x \rightarrow f(x)$ with a certain domain $D(f)$ and range $R(f)$, namely, the *ordering* of $D(f)$. The pages of a book, and pictures of a film are numbered consecutively, and the domain of a directory is also ordered alphabetically. In the case of a book or film the ordering helps to make sense out of the material, and a dictionary without any order is almost useless. Of course, swapping through films has become a part of the life-style of to-day, but the risk of a loss of understanding is obvious. To be able to catch the main

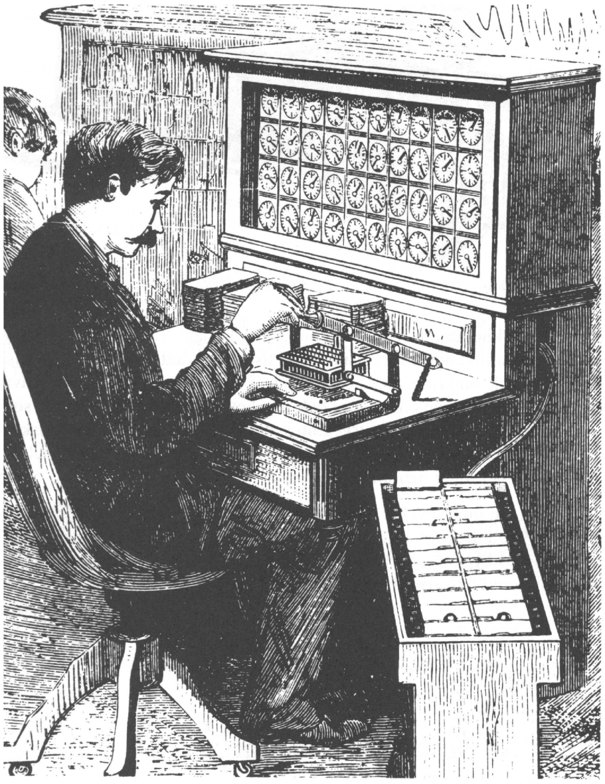


FIGURE 9.3. Hermann Hollerith, inventor of the punched card machine: "My friend Dr. Billings one night at the pub suggested to me that there ought to be some mechanical way of doing the census, something on the principle of the Jacquard loom, whereby holes in a card regulate the pattern to be woven".

idea or plot of a book or film *as a whole* it is necessary to read the pages or view the pictures in order. The ordering helps us to get an overall meaning.

Similarly, it is useful to be able to catch the main properties of a function $f(x)$, and this can sometimes be done by graphing or visualizing the function using some suitable ordering of $D(f)$. We now go into the topic of graphing functions $f(x)$ with $D(f)$ and $R(f)$ subsets of $\mathbb{Q}$, of course with the usual ordering of $\mathbb{Q}$, and with the purpose of trying to grasp the nature of a given function "as a whole".

9.3 Graphing functions of integers

So far we have described a function both by listing all its values in a table like the phonebook and by giving a formula like $f(n) = n^2$ and indicating the domain. It is also useful to have a picture of the behavior of a function, or in other words, to represent a function geometrically. Graphing functions is a way of visualizing a function so that we can grasp the nature of the function "in one shot" or as one object. For example, we can describe the function as increasing in this region and decreasing in this other region, giving an idea of how it behaves without being specific.

We begin by describing the graphing of functions $f : \mathbb{Z} \to \mathbb{Z}$. Recall that integers are represented geometrically using the integer line. To describe the input and output to a function $f : \mathbb{Z} \to \mathbb{Z}$, we therefore need two number lines so that we can mark the points in $D(f)$ on one and the points in $R(f)$ on the other. A convenient way to arrange these two number lines is to place them orthogonal to each other as in Fig. 9.4. If we mark the points obtained by intersecting vertical lines through integer points on the horizontal axis with the horizontal lines through integer points on the vertical axis, we get a grid of points like that shown in Fig. 9.4. This is called the *integer coordinate plane*. Each number line is called an *axis* of the coordinate plane

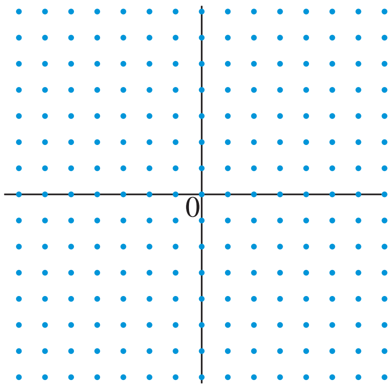


FIGURE 9.4. The integer coordinate plane.

while the intersection point of the two number lines is called the *origin* and is denoted by 0.

As we saw, a function $f : \mathbb{Z} \to \mathbb{Z}$ can be represented by making a list with the inputs placed side-by-side with the corresponding outputs. We show such a table for $f(n) = n^2$ in Fig. 9.5. We can represent such a table also in the integer coordinate plane by marking only those points corresponding to an entry in the table, i.e. marking each intersection point of the line rising vertically from the input and the line extending horizontally from the corresponding output. We draw the plot corresponding to $f(n) = n^2$ in Fig. 9.6.

EXAMPLE 9.11. In Fig. 9.7, we plot n, n^2, and 2^n along the vertical axis with $n = 1, 2, 3, .., 6$ along the horizontal axis. The plot suggests

n	$f(n)$
0	0
1	1
-1	1
2	4
-2	4
3	9
-3	9

FIGURE 9.5. A tabular listing of $f(n) = n^2$.

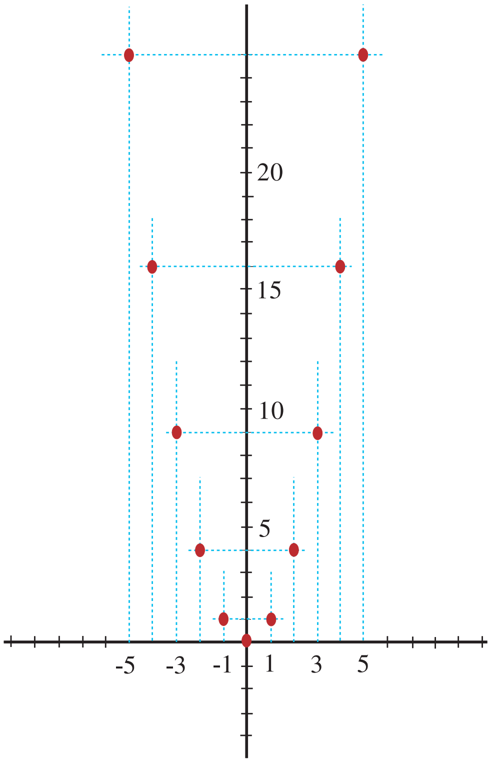


FIGURE 9.6. Points associated with the function $f(n) = n^2$ with $D(f) = \mathbb{Z}$.

2^n grows more quickly than both n and n^2 as n increases. In Fig. 9.8,

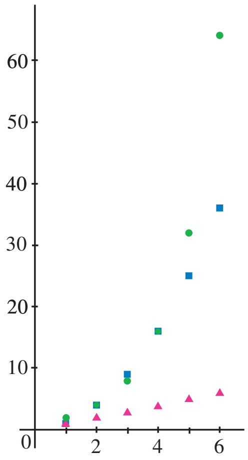


FIGURE 9.7. Plots of the functions ▲ $f(n) = n$, ■ $f(n) = n^2$, and ♦ $f(n) = 2^n$ with $D(f) = \mathbb{N}$.

we plot n^{-1}, n^{-2}, and 2^{-n} with $n = 1, 2, .., 6$, and we see that 2^{-n} decreases most rapidly and n^{-1} least rapidly. Compare these results to Fig. 9.7.

Instead of using a table to list the points for a function, we can represent a point on the integer plane mathematically by means of an *ordered pair* of numbers. To the point in the plane located at the intersection of the vertical line passing through n on the horizontal axis and the horizontal line passing through m on the vertical axis, we associate the pair of numbers (n, m). These are the *coordinates* of the point. Using this notation, we can describe the function $f(n) = n^2$ as the set of ordered pairs

$$\{(0,0), (1,1), (-1,1), (2,4), (-2,4), (3,9), (-3,9), \cdots\}.$$

Note that we always associate the first number in the ordered pair with the horizontal location of the point and the second number with the vertical location. This is an arbitrary choice.

We can illustrate the idea of a function giving a transformation of its domain into its range nicely using its graph. Consider Fig. 9.6. We start at a point in the domain on the horizontal axis and follow a line straight up to the point on the graph of the function. From this point, we follow a line horizontally to the vertical axis. In other words, we can find the output associated to a given input by tracing first a vertical line and then a horizontal line.

Note also that for functions with $D(f) = \mathbb{N}$ or $D(f) = \mathbb{Q}$ it is only possible to graph part of the function, simply because we cannot in practice

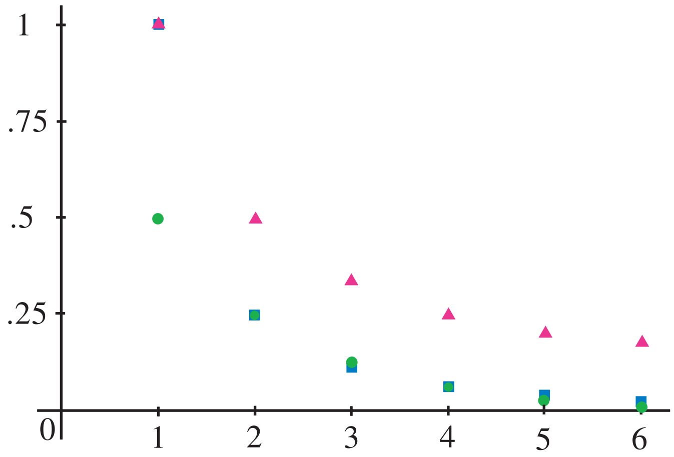


FIGURE 9.8. Plots of the functions ▲ $f(n) = n^{-1}$, ■ $f(n) = n^{-2}$, and ♦ $f(n) = 2^{-n}$ with $D(f) = \mathbb{N}$.

extend the natural or integer number line all the way to "infinity". Of course, a table representation of such a function must also be limited to a finite range of argument values. Only a defining formula of the function values, like $f(n) = n^2$ (together with a specification of $D(f)$), can give the full picture in this case.

9.4 Graphing functions of rational numbers

Now we consider plotting a function $f : \mathbb{Q} \to \mathbb{Q}$. Following the lead of functions of integers, we plot functions of rational numbers on the *rational coordinate plane* which we construct by placing two rational number lines called the axes at right angles and meeting at the origins and then marking every point that has rational number coordinates. Of course considering Fig. 7.4, such a plane will appear to be solid even if it is not solid. We avoid plotting an example!

If we begin the plot of a function of rational numbers as above by writing down a list of values, we realize immediately that graphing a function of rational numbers is more complicated than graphing a function of integers. When we compute values of a function of integers, we cannot compute *all* the values because there are infinitely many integers. Instead we choose a smallest and largest integer and compute the values of the functions for those integers in between. For the same reason, we can not compute all the values of a function defined on the rational numbers. But now we have to cut off the list also in another way: we have to choose a smallest and largest number for making the list as before, but we also have to decide how many points to use in between the low and high values. In other words, we cannot compute the values of the function at *all* the rational numbers in between two rational numbers. This means that a list of values of a function of rational numbers always has "gaps" in between the points where we evaluate the function. We give an example to make this clear.

EXAMPLE 9.12. We list some values of the function $f(x) = \frac{1}{2}x + \frac{1}{2}$ defined on the rational numbers:

x	$\frac{1}{2}x+\frac{1}{2}$
-5	-2
-2.8	$-.9$
-2	$-.5$
-1.2	$-.1$
-1	0

x	$\frac{1}{2}x+\frac{1}{2}$
$-.6$	$.2$
$.2$	$.6$
1	1
3	2
5	3

and then plot the function values in Fig. 9.9.

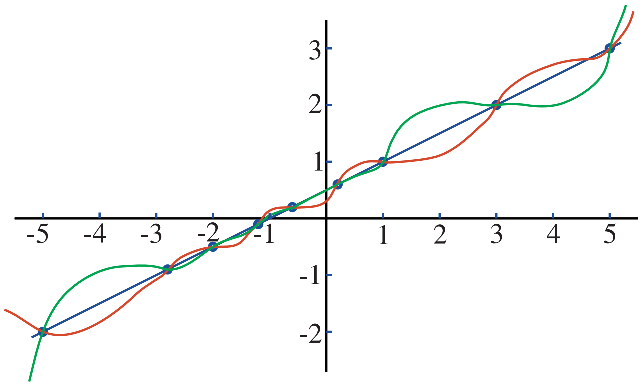


FIGURE 9.9. A plot of the function values of $f(x) = \frac{1}{2}x+\frac{1}{2}$, and several functions taking on the same values at the sample points.

The values we list for this example suggest strongly that we should draw a straight line through the indicated points in order to plot the function. However, we cannot be sure that this is the correct graph because there are many functions that agree with $\frac{1}{2}x + \frac{1}{2}$ at the points we computed, for example we show two of them in Fig. 9.9. Therefore, to graph a function accurately, we would need to evaluate it at many more points than we have used in Fig. 9.9 in general. On the other hand we cannot possibly compute the values $f(x)$ for all possible rational numbers x, so that in the end we still have to guess the values of the function in between the points we compute, assuming that the function does not do anything strange there. Matlab for example fills the gaps between the computed points with straight line segments when plotting.

Deciding whether or not we have evaluated a function defined on the rational numbers enough times to be able to guess its behavior is an interesting and important problem. This is not just a theoretical problem by the way: if we have to measure some quantities during an experiment that should theoretically lie on a line, we are very likely to get a plot of a function that is close to a line, but that has little wiggles because of experimental error.

In fact we are able to use Calculus to help with this decision. For now, we will assume that the functions we plot vary smoothly between the sample points, which is largely true for the functions we consider in this book.

We finish this chapter by giving another example of a plot. In the next chapter, we spend a lot more time on plotting.

EXAMPLE 9.13. We list some values of the function $f(x) = x^2$ defined on the rational numbers:

x	x^2	x	x^2	x	x^2
-4	16	$-.8$	.64	2.3	5.29
-3.5	12.25	$-.4$	.16	2.4	5.76
-3.1	9.618	0	0	3	9
-2	4	.2	.04	3.1	9.61
-1.8	3.24	1.2	1.44	3.6	12.96
-1.4	1.96	1.5	2.25	3.7	13.69
-1	1	2.21	4.8841	4	16

and then plot the function values in Fig. 9.10.

9.5 A function of two variables

We give an example of a function of two variables. The total cost in the Dinner Soup/Ice Cream model was

$$15x + 3y,$$

where x was the amount of beef and y that of ice cream. We may view the total cost $15x + 3y$ as a function $f(x, y) = 15x + 3y$ of the two variables x and y. For each value of x and y there is a corresponding function value $f(x, y) = 15x + 3y$ representing the total cost. We think here of both x and y as independent variables which may vary freely, corresponding to any combination of beef and ice cream, and the function value $z = f(x, y)$ as a dependent variable. For each pair of values of x and y there is assigned a value of $z = f(x, y) = 15x + 3y$. We may write $(x, y) \rightarrow f(x, y) = 15x + 3y$, denoting the pair of x and y by (x, y).

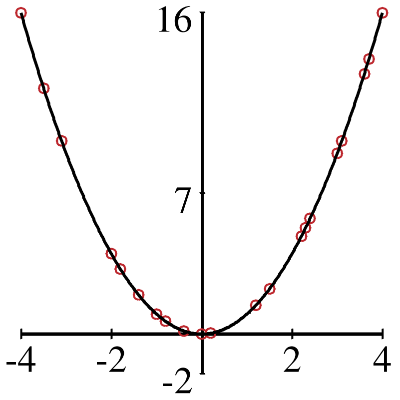


FIGURE 9.10. A plot of some of the points given by $f(x) = x^2$ and a smooth curve that passes through the points.

This represents a very natural and very important extension of the concept of a function considered so far: a function may depend on two independent variables. Assuming that for the function $f(x, y) = 15x + 3y$ we allow both x and y to vary over $[0, \infty)$, we will write $f : [0, \infty) \times [0, \infty) \to [0, \infty)$ to denote that for each $x \in [0, \infty)$ and $y \in [0, \infty)$, that is for each pair $(x, y) \in [0, \infty) \times [0, \infty)$, there is a unique value $f(x, y) = 15x + 3y \in [0, \infty)$ assigned.

EXAMPLE 9.14. The prize your roommate has to pay for the x pounds of beef and y pounds of ice cream is $p = 15x + 3y$, that is $p = f(x, y)$ where $f(x, y) = 15x + 3y$.

EXAMPLE 9.15. The time t required for a certain bike trip depends on the distance s of the trip, and on the (mean) speed v as $\frac{s}{v}$, that is $t = f(s, v) = \frac{s}{v}$.

EXAMPLE 9.16. The pressure p in an ideal (thin) gas mixture depends on the temperature T and volume V occupied of the gas as $p = f(T, V) = \frac{nRT}{V}$, where n is the number of moles of gas molecules and R is the universal gas constant.

9.6 Functions of several variables

Of course we may go further and consider functions depending on several independent variables.

EXAMPLE 9.17. Letting your roommate decide the amount of beef x, carrots y and potatoes z in the Dinner Soup, the cost k of the soup will be $k = 8x + 2y + z$ depending on the three variables x, y and z. The cost is thus given by $k = f(x, y, z)$ where $f(x, y, z) = 8x + 2y + z$.

EXAMPLE 9.18. The temperature u at a certain position depends on the three space coordinates x, y and z, as well as on time t, that is $u = u(x, y, z, t)$.

As we come to consider situations with more than just a few independent variables, it soon becomes necessary to change notation and use some kind of indexation of the variables like for example denoting the spacial coordinates x, y and z instead by x_1, x_2 and x_3. For example, we may then write the function u in the last example as $u(x, t)$ where $x = (x_1, x_2, x_3)$ contains the three space coordinates.

Chapter 9 Problems

9.1. Identify four functions you encounter in your daily life and determine the domain and range for each.

9.2. For the function $f(x) = 4x - 2$, determine the range corresponding to: (a) $D(f) = (-2, 4]$, (b) $D(f) = (3, \infty)$, (c) $D(f) = \{-3, 2, 6, 8\}$.

9.3. Given that $f(x) = 2 - 13x$, find the domain $D(f)$ corresponding to the range $R(f) = [-1, 1] \cup (2, \infty)$.

9.4. Determine the domain and range of $f(x) = x^3/100 + 75$ where $f(x)$ is a function giving the temperature inside an elevator holding x people and with a maximum capacity of 9 people.

9.5. Determine the domain and range of $H(t) = 50 - t^2$ where $H(t)$ is a function giving the height in meters of a ball dropped at time $t = 0$.

9.6. Find the range of the function $f(n) = 1/n^2$ defined on $D(f) = \{n \in \mathbb{N} : n \geq 1\}$.

9.7. Find the domain and a set B containing the range of the function $f(x) = 1/(1 + x^2)$.

9.8. Find the domain of the functions

$$\text{(a)}\ \frac{2-x}{(x+2)x(x-4)(x-5)} \qquad \text{(b)}\ \frac{x}{4-x^2} \qquad \text{(c)}\ \frac{1}{2x+1} + \frac{x^2}{x-8}$$

9.9. *(Harder)* Consider the function $f(n)$ defined on the natural numbers where $f(n)$ is the remainder obtained by dividing n by 5 using long division. So for example, $f(1) = 1$, $f(6) = 1$, $f(12) = 2$, etc. Determine $R(f)$.

9.10. Illustrate the map $f : \mathbb{N} \rightarrow \mathbb{Q}$ using two intervals where $f(n) = 2^{-n}$.

9.11. Plot the following functions $f : \mathbb{Z} \rightarrow \mathbb{Z}$ after making a list of at least 5 values: (a) $f(n) = 4 - n$, (b) $f(n) = 2n - n^2$, (c) $f(n) = (n+1)^3$.

9.12. Draw three different curves that pass through the points $(-2,-1)$, $(-1,-.5)$, $(0,.25)$, $(1,1.5)$, $(3,4)$.

9.13. Plot the functions; (a) 2^{-n}, (b) 5^{-n}, and (c) 10^{-n}; defined on the natural numbers n. Compare the plots.

9.14. Plot the function $f(n) = \frac{10}{9}(1 - 10^{-n-1})$ defined on the natural numbers.

9.15. Plot the function $f : \mathbb{Q} \rightarrow \mathbb{Q}$ with $f(x) = x^3$ after making a table of values.

9.16. Write a *MATLAB*© function that takes two rational arguments x and y and returns their sum $x + y$.

9.17. Write a *MATLAB*© function that takes two arguments x and y representing two velocities, and returns the time gained per kilometer by raising the velocity from x to y.

10
Polynomial functions

Sometimes he thought to himself, “Why?” and sometimes he thought, “Wherefore?”, and sometimes he thought, “Inasmuch as which?”. (Winnie-the Pooh)

He was one of the most original and independent of men and never did anything or expressed himself like anybody else. The result was that it was very difficult to take notes at his lectures so that we had to trust mainly to Rankine’s text books. Occasionally in the higher classes he would forget all about having to lecture and, after waiting for ten minutes or so, we sent the janitor to tell him that the class was waiting. He would come rushing into the door, taking a volume of Rankine from the table, open it apparently at random, see some formula or other and say it was wrong. He then went up to the blackboard to prove this. He wrote on the board with his back to us, talking to himself, and every now and then rubbed it all out and said it was wrong. He would then start afresh on a new line, and so on. Generally, towards the end of the lecture he would finish one which he did not rub out and say that this proved Rankine was right after all. (Rayleigh about Reynolds)

10.1 Introduction

We now proceed to study polynomial functions, which are fundamental in Calculus and Linear Algebra. A *polynomial function*, or *polynomial*, $f(x)$

has the form

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots + a_n x^n, \tag{10.1}$$

where a_0, a_1, $\cdots$, a_n, are given rational numbers called the *coefficients* and the variable x varies over some set of rational numbers. The value of a polynomial function $f(x)$ can be directly computed by adding and multiplying rational numbers. The Dinner Soup function $f(x) = 15x$ is an example of a *linear polynomial* with $n = 1$, $a_0 = 0$ and $a_1 = 15$, and the Muddy Yard function $f(x) = x^2$ is an example of a *quadratic function* with $n = 2$, $a_0 = a_1 = 0$, and $a_2 = 1$.

If all the coefficients a_i are zero, then $f(x) = 0$ for all x and we say that $f(x)$ is the *zero polynomial.* If n denotes the largest subscript with $a_n \neq 0$, we say that the *degree* of $f(x)$ is n. The simplest polynomials besides the zero polynomial are the *constant* polynomials $f(x) = a_0$ of degree 0. The next simplest cases are the *linear* polynomials $f(x) = a_0 + a_1 x$ of degree 1 and *quadratic* polynomials $f(x) = a_0 + a_1 x + a_2 x^2$ of degree 2 (assuming $a_1 \neq 0$ respectively $a_2 \neq 0$), which we just gave examples of, and we met a polynomial of degree 3 in the model of solubility of $Ba(IO_3)_2$ in Proposition 7.10.

The polynomials are basic "building blocks" in the mathematics of functions, and spending some effort understanding polynomials and learning some facts about them will be very useful to us later on. Below we will meet other functions such as the *elementary functions* including trigonometric functions like $\sin(x)$ and the exponential function $\exp(x)$. The elementary functions are all solutions of certain fundamental differential equations, and evaluation of these functions requires solution of the corresponding differential equation. Thus, these functions are not called elementary because they are elementary to evaluate, like a polynomial, but because they satisfy fundamental "elementary" differential equations.

In the history of mathematics, there has been two grand attempts to describe "general functions" in terms of (i) polynomial functions (power series) or (ii) trigonometric functions (Fourier series). In the *finite element method* of our time, general functions are described using *piecewise polynomials.*

We start with linear and quadratic functions, before considering general polynomial functions.

10.2 Linear polynomials

We start with the linear polynomial $y = f(x) = mx$, where m is a rational number. We write here m instead of a_1 because this notation is often used. We may choose $D(p) = \mathbb{Q}$ and if $m \neq 0$ then $R(p) = \mathbb{Q}$ because if y is any rational number, then $x = y/m$ inserted into $f(x) = mx$ gives the value of

$f(x) = y$. In other words the function $f(x) = mx$ with $m \neq 0$ maps $\mathbb{Q}$ onto $\mathbb{Q}$.

One way to view the set of (x, y) that satisfy $y = mx$ is to realize that such (x, y) also satisfy $y/x = m$. Suppose that (x_0, y_0) and (x_1, y_1) are two points satisfying $y/x = m$. If we draw a triangle with one corner at the origin and with one side parallel to the x axis of length x_0 and another side parallel to the y axis of length y_0 then draw the corresponding triangle for the other point with sides of length x_1 and y_1, see Fig. 10.1, then the condition

$$\frac{y_0}{x_0} = m = \frac{y_1}{x_1}$$

means that these two triangles are similar. In fact any point (x, y) satisfying $y/x = m$ must form a triangle similar to the triangle made by (x_0, y_0), see Fig. 10.1. This means that such points lie on a line that passes through the origin as indicated.

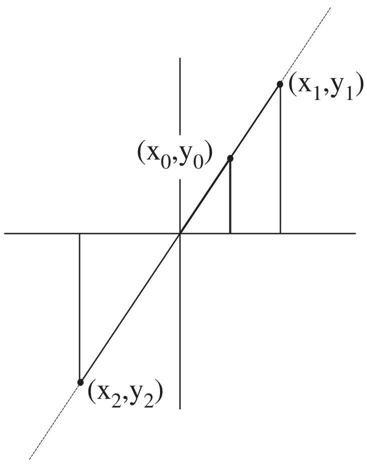


FIGURE 10.1. Points satisfying $y = mx$ form similar triangles. In this figure, $m = 3/2$.

In the language of architecture, m, or the ratio of y to x, is called the *rise over the run* while mathematicians call m the *slope* of the line. If we imagine standing on a straight road going up hill, then the slope tells how much we have to climb for any horizontal distance we travel. In other words, the larger the slope m, the steeper the line. By the way, if the slope is negative then the line slopes downwards. We show some different lines in Fig. 10.2. When the slope $m = 0$, then we get a horizontal line sitting on top of the x axis. A vertical line on the other hand is the set of points (x, y) where $x = a$ for some constant a. Vertical lines do not have a well defined slope.

Using the slope to describe how a line increases or decreases does not depend on the line passing through the origin. We can start at any point on

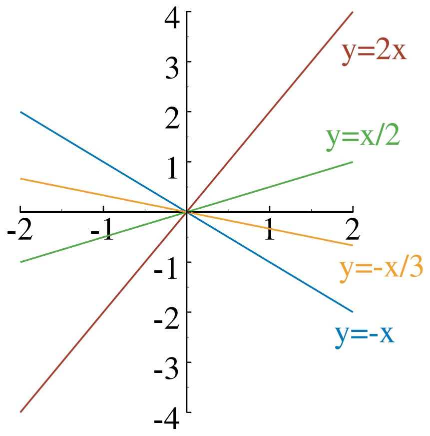


FIGURE 10.2. Examples of lines.

a line and ask how much the line rises or lowers if we move horizontally a distance x, see Fig. 10.3. If the points are (x_0, y_0) and (x_1, y_1), then $y_1 - y_0$

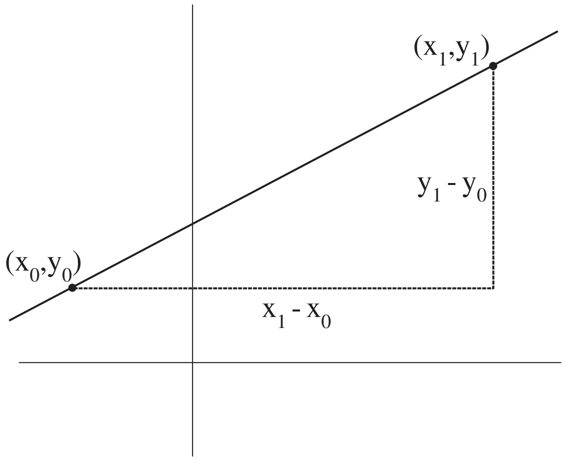


FIGURE 10.3. The slope of any line is determined by the amount of rise over the amount of run between any two points on the line.

is the amount of "rise" corresponding to the "run" of $x_1 - x_0$. Hence the slope of the line through the points (x_0, y_0) and (x_1, y_1) is

$$m = \frac{y_1 - y_0}{x_1 - x_0}.$$

If (x, y) is any other point on the line, then we know that

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

or

$$(y - y_0) = m(x - x_0). \tag{10.2}$$

This is called the *point-slope* equation for a line.

EXAMPLE 10.1. We find the equation of the line through $(4, -5)$ and $(2, 3)$. The slope is

$$m = \frac{3 - (-5)}{2 - 4} = 4$$

and the line is $y - 3 = 4(x - 2)$.

We can rewrite (10.2) to resemble (10.1) by multiplying out the terms in (10.2) and solving for y. This yields the *slope-intercept* form:

$$y = mx + b, \tag{10.3}$$

with $b = y_1 - mx_1$. b is called the y*-intercept* of the line because the line crosses the y axis at the point $(0, b)$, i.e. at a height of b. The difference between the graphs of $y = mx$ and $y = mx + b$ is simply that every point on $y = mx + b$ is *translated* vertically a distance of b from the corresponding point on $y = mx$. In other words, we can graph $y = mx + b$ by first graphing $y = mx$ and the moving the line vertically by an amount of b. We illustrate in Fig. 10.4. When $b > 0$ we move the line up and when $b < 0$ we move the

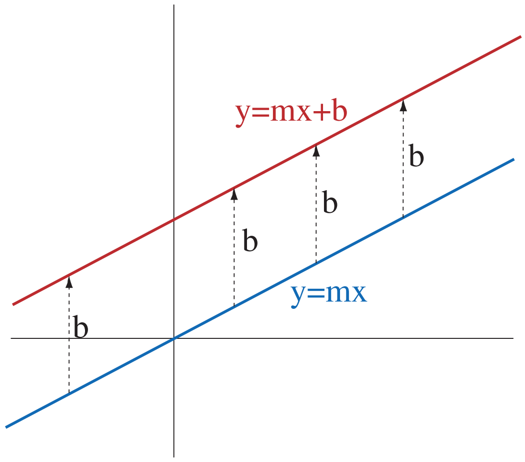


FIGURE 10.4. The graph of $y = mx + b$ is found by translating the graph of $y = mx$ vertically by an amount b. In this case $b > 0$.

line down.

We can find the slope-intercept form directly from knowing two points.

EXAMPLE 10.2. We find the slope-intercept equation of the line through $(-3,5)$ and $(4,1)$. The slope is

$$m = \frac{5-1}{-3-4} = -\frac{4}{7}.$$

To compute the y-intercept, we substitute either point into the equation $y = -\frac{4}{7}x + b$, for example,

$$5 = -\frac{4}{7} \times -3 + b$$

so $b = 23/7$ and $y = -\frac{4}{7}x + \frac{23}{7}$.

The technique of translating a known graph can be very useful when graphing. For example once we have plotted $y = 4x$, we can quickly plot $y = 4x - 12$, $y = 4x - \frac{1}{5}$, $y = 4x + 1$, and $y = 4x + 113.45$ by translation.

10.3 Parallel lines

We now draw a connection to the parallel axiom of Euclidean geometry, which we discussed in chapter Euclid and Pythagoras. First, let $y = mx+b_1$ and $y = mx + b_2$ be two lines with same slope m, but different y-intercepts b_1 and b_2, so that the lines are not identical. These two lines cannot ever cross, since there is no x for which $mx + b_1 = mx + b_2$, because $b_1 \neq b_2$. We conclude that two lines with the same slope are parallel in the sense of Euclidean geometry.

On the other hand, if $y = m_1x + b_1$ and $y = m_2x + b_2$ are two lines with different slopes $m_1 \neq m_2$, then the two lines will cross, since we can solve the equation $m_1x+b_1 = m_2x+b_2$ uniquely, to get $x = (b_1 - b_2)/(m_2 - m_1)$. We conclude that two lines corresponding to two linear polynomials $y = m_1x + b_1$ and $y = m_2x + b_2$ are parallel if and only if $m_1 = m_2$.

EXAMPLE 10.3. We find the equation of the line that is parallel to the line through $(2,5)$ and $(-11,6)$ and passing through $(1,1)$. The slope of the line must be $m = (6-5)/(-11-2) = -1/13$. Therefore, $1 = -1/13 \times 1 + b$ or $b = 14/13$ and $y = -\frac{1}{13}x + \frac{14}{13}$.

EXAMPLE 10.4. We can find the point of intersection between the line $y = 2x+3$ and $y = -7x-4$ by setting $2x+3 = -7x-4$. Adding $7x$ and subtracting 3 from both sides $2x + 3 + 7x - 3 = -7x - 4 + 7x - 3$ gives $9x = -7$ or $x = -7/9$. We can get the value of y from either equation, $y = 2x + 3 = 2(\frac{-7}{9}) + 3 = \frac{13}{9}$ or $y = -7x - 4 = -7(\frac{-7}{9}) - 4 = \frac{13}{9}$.

10.4 Orthogonal lines

Lets us next show that two lines corresponding to two linear polynomials $y = m_1x + b_1$ and $y = m_2x + b_2$ are *orthogonal*, that is make an angle of $90°$ or $270°$, if and only if $m_1m_2 = -1$.

Since the values of b_1 and b_2 can be changed without changing the directions of the lines, it is sufficient to show that the statement is true for two lines that pass through the origin. Assume now that the lines are orthogonal. Then m_1 and m_2 must have different signs, since otherwise either both of the lines are increasing or both are decreasing and then they cannot be perpendicular. Now consider the triangles drawn in Fig. 10.5. The lines are

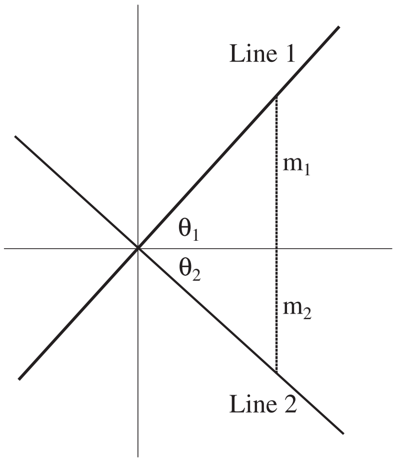


FIGURE 10.5. Similar triangles defined by perpendicular lines with slope m_1 and m_2. The angles θ_1 and θ_2 add up to $90°$.

perpendicular only if the angles θ_1 and θ_2 that the lines make with the x axis add up to $90°$. This can happen only if the triangles drawn are similar. This means that $1/|m_1| = 1/|m_2|$ or $|m_1|\,|m_2| = 1$. This shows the result since m_1 and m_2 have opposite signs or $m_1m_2 < 0$.

Finally, assuming that $m_1m_2 = -1$ shows that the two triangles are similar and the orthogonality follows.

EXAMPLE 10.5. We find the equation of the line that is perpendicular to the line through $(2, 5)$ and $(-11, 6)$ and passing through $(1, 1)$. The slope of the first line is $m = (6 - 5)/(-11 - 2) = -1/13$, so the slope of the line we compute is $-1/(-1/13) = 13$. Therefore, $1 = 13 \times 1 + b$ or $b = -12$ and $y = 13x - 12$.

We will return to the topic of parallel and orthogonal lines in a little wider setting in chapter Analytic geometry in $\mathbb{Q}^2$. In particular, the so far excluded cases with vertical or horisontal lines, will then be included in a natural way.

10.5 Quadratic polynomials

The general quadratic polynomial has the form

$$f(x) = a_2x^2 + a_1x + a_0$$

for constants a_2, a_1, and a_0, where we assume $a_2 \neq 0$ (otherwise we go back to the linear case).

We show how to plot such a function by using the idea of plotting lines in the previous section starting with the simplest example of a quadratic function

$$y = f(x) = x^2.$$

The domain of f is the set of rational numbers while the range contains some of the nonnegative rational numbers. We list some of the values here:

x	x^2
-3	9
-2	4
-1	1
$-.5$	.25
$-.25$	.125
$-.1$	.01
0	0

x	x^2
.1	.01
.25	.125
.5	.25
1	1
2	4
3	9
4	16

We also observe that $f(x) = x^2$ is *increasing* for $x > 0$, which means that if $0 < x_1 < x_2$ then $f(x_1) < f(x_2)$. This follows because $x_1 < x_2$ means that $x_1 \times x_1 < x_2 \times x_1 < x_2 \times x_2$. Likewise, we can show that $f(x) = x^2$ is *decreasing* for $x < 0$, which means that if $x_1 < x_2 < 0$ then $f(x_1) > f(x_2)$. This means that the function at least cannot wiggle very much in between the values we compute. We plot the values of $f(x) = x^2$ in Fig. 10.6 for 601 equally spaced points between $x = -3$ and $x = 3$.

To draw the graph of a general quadratic function, we follow the idea behind computing the graphs of lines by using translation. We start with $f(x) = x^2$ and then change that graph to get the graph of any other quadratic. There are two kinds of changes we make.

The first change is called *scaling*. Consider the plots of the quadratic functions in Fig. 10.7. Each of these functions has the form $y = f(x) = a_2x^2$ for a constant a_2. Their plots all have the same basic shape as $y = x^2$. However the heights of the points on $y = a_2x^2$ are a factor of $|a_2|$ higher or lower than the height of the corresponding point on $y = x^2$: higher if $|a_2| > 1$ and lower if $|a_2| < 1$. If $a_2 < 0$ then the plot is also "flipped" or *reflected* through the x-axis.

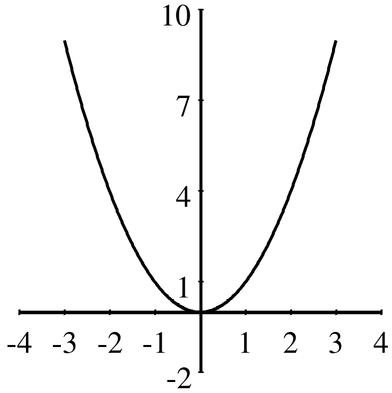


FIGURE 10.6. Plot of $f(x) = x^2$. The function is decreasing for $x < 0$ and increasing for $x > 0$.

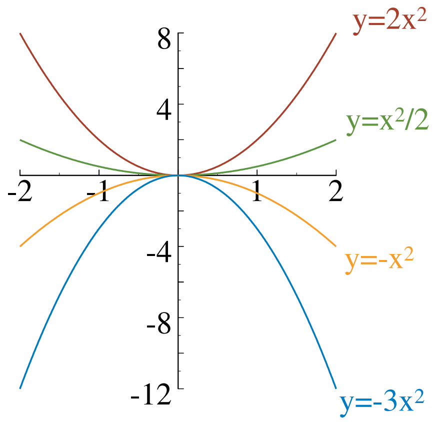


FIGURE 10.7. Plots of $y = x^2$ scaled four different ways.

The second change we consider is translation. The two possibilities are to translate horizontally, or sideways, and vertically. We show examples of both in Fig. 10.8. Graphs of quadratic functions of the form $f(x) = (x+x_0)^2$ can be drawn by moving the graph of $y = x^2$ sideways to the right a distance of $|x_0|$ if $x_0 < 0$ and to the left a distance of x_0 if $x_0 > 0$. The easiest way to remember which direction to translate is to figure out the new position of the *vertex*, which is the lowest or highest point of the quadratic. For $y = (x-1)^2$, the lowest point is $x = 1$ and the graph is obtained by moving the graph of $y = x^2$ so the vertex is now at $x = 1$. For $y = (x + .5)^2$, the vertex is at $x = -.5$ and we get the graph by moving the graph of $y = x^2$ to the left a distance of .5. On the other hand, the graph of a function $y = x^2 + d$ can be obtained by translating the graph of $y = x^2$ vertically, in a fashion similar to what we did for lines. Recall that $d > 0$ translates the graph upwards and $d < 0$ downwards.

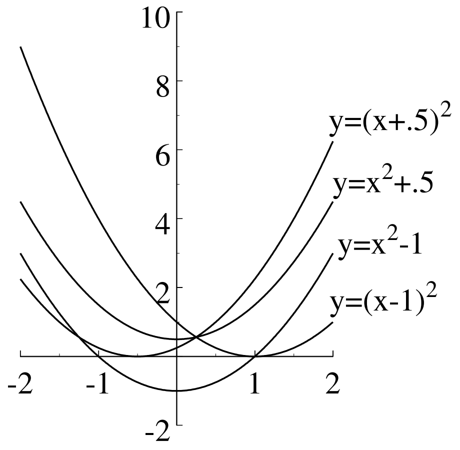


FIGURE 10.8. Plots of $y = x^2$ translated four different ways.

Now it is possible to put all of this together to plot the graph of the function $y = f(x) = a(x - x_0)^2 + d$ by scaling and translating the graph of $y = x^2$. We perform each operation in the same order that we would use to do the arithmetic in computing values of $f(x)$; first translate horizontally by x_0, then scale by a, and finally translate vertically by d.

EXAMPLE 10.6. We plot $y = -2(x + 1)^2 + 3$ in Fig. 10.9 by starting with $y = x^2$ in (a), translating horizontally to get $y = (x + 1)^2$ in (b), scaling vertically to get $y = -2(x + 1)^2$ in (c), and finally translating vertically to get $y = -2(x + 1)^2 + 3$ in (d).

The last step is to consider the plot of the quadratic $y = ax^2 + bx + c$. The idea is to first rewrite this in the form $y = a(x - x_0)^2 + d$ for some x_0

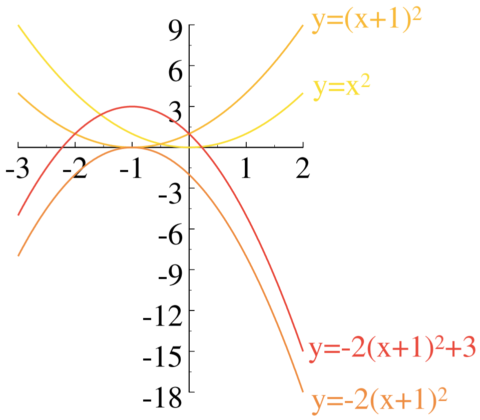


FIGURE 10.9. Plotting $y = -2(x+1)^2 + 3$ in a systematic way.

and d, then we can draw the graph easily. To explain how to do this, we work backwards using the example $y = -2(x+1)^2 + 3$. Multiplying out, we get

$$y = -2(x^2 + 2x + 1) + 3 = -2x^2 - 4x - 2 + 3 = -2x^2 - 4x + 1.$$

Now if we are given $y = -2x^2 - 4x + 1$, we can do the following steps

$$\begin{aligned} y &= -2x^2 - 4x + 1 \\ &= -2(x^2 + 2x) + 1 \\ &= -2(x^2 + 2x + 1 - 1) + 1 \\ &= -2(x^2 + 2x + 1) + 2 + 1 \\ &= -2(x+1)^2 + 3. \end{aligned}$$

This procedure is called *completing the square*. Given $x^2 + bx$, the idea to add the number m so that $x^2 + bx + m$ is the square $(x - x_0)^2$ for some appropriate x_0. Of course we also have to subtract m so we don't change the function. *Note that we added and subtracted* 1 *inside the parenthesis in the example above!* If we multiply out, we get

$$(x - x_0)^2 = x^2 - 2x_0 x + x_0^2$$

which is supposed to match

$$x^2 + bx + m.$$

This means that $x_0 = -b/2$ while $m = x_0^2 = b^2/4$. In the example above, $b = 2$, $x_0 = -1$, and $m = 1$.

EXAMPLE 10.7. We complete the square on $y^2 - 3x + 7$. Here $b = -3$, $x_0 = 3/2$, and $m = 9/4$. So we write

$$\begin{aligned} y^2 - 3x + 7 &= y^2 - 3x + \frac{9}{4} - \frac{9}{4} + 7 \\ &= \left(y - \frac{3}{2}\right)^2 + \frac{19}{4}. \end{aligned}$$

EXAMPLE 10.8. We complete the square on $6y^2 + 4y - 2$. We first have to write

$$6y^2 + 4y - 2 = 6\left(y^2 + \frac{2}{3}y\right) - 2.$$

Now $b = 2/3$, $x_0 = -1/3$, and $m = 1/9$. So we write

$$\begin{aligned} 6y^2 + 4y - 2 &= 6\left(y^2 + \frac{2}{3}y + \frac{1}{9} - \frac{1}{9}\right) - 2 \\ &= 6\left(y + \frac{1}{3}\right)^2 - \frac{6}{9} - 2 \\ &= 6\left(y + \frac{1}{3}\right)^2 - \frac{8}{3}. \end{aligned}$$

EXAMPLE 10.9. We complete the square on $y = \frac{1}{2}x^2 - 2x + 3$.

$$\begin{aligned} \frac{1}{2}x^2 - 2x + 3 &= \frac{1}{2}(x^2 - 4x) + 3 \\ &= \frac{1}{2}(x^2 - 4x + 4 - 4) + 3 \\ &= \frac{1}{2}(x - 2)^2 - 2 + 3 \\ &= \frac{1}{2}(x - 2)^2 + 1. \end{aligned}$$

10.6 Arithmetic with polynomials

We turn now to investigating properties of polynomials of general degree, beginning with arithmetic properties. Recall that if we add, subtract, or multiply two rational numbers, then the result is another rational number. In this section, we show that the analogous property holds for polynomials.

10.6.1 The Σ notation for finite sums

Before exploring arithmetic with polynomials, we introduce a convenient notation for dealing with long finite sums using the Greek letter sigma Σ.

Given any $n+1$ quantities $\{a_0, a_1, \cdots, a_n\}$ indexed with subscripts, we write the sum

$$a_0 + a_1 + \cdots + a_n = \sum_{i=0}^{n} a_i.$$

The *index* of the sum is i and it is assumed that it takes on all the integers between the *lower limit*, which is 0 here, and the *upper limit*, which is n here, of the sum.

EXAMPLE 10.10. The finite *harmonic series* of order n is

$$\sum_{i=1}^{n} \frac{1}{i} = 1 + \frac{1}{2} + \frac{1}{3} + \cdots \frac{1}{n}$$

while the finite *geometric series* of order n with factor r is

$$1 + r + r^2 + \cdots + r^n = \sum_{i=0}^{n} r^i.$$

Notice that the index i is a *dummy variable* in the sense that it can be renamed or the sum can be rewritten to start at another integer.

EXAMPLE 10.11. The following sums are all the same:

$$\sum_{i=1}^{n} \frac{1}{i} = \sum_{z=1}^{n} \frac{1}{z} = \sum_{i=0}^{n-1} \frac{1}{i+1} = \sum_{i=4}^{n+3} \frac{1}{i-3}.$$

Using the Σ notation, we can write the general polynomial (10.1) in the more condensed form:

$$f(x) = \sum_{i=0}^{n} a_i x^i = a_0 + a_1 x^1 + \cdots + a_n x^n.$$

EXAMPLE 10.12. We can write

$$1 + 2x + 4x^2 + 8x^3 + \cdots + 2^{20}x^{20} = \sum_{i=0}^{20} 2^i x^i$$

and

$$1 - x + x^2 - x^3 - \cdots - x^{99} = \sum_{i=0}^{99} (-1)^i x^i.$$

since $(-1)^i = 1$ if i is even and $(-1)^i = -1$ if i is odd.

10.6.2 Addition of polynomials

Given two polynomials

$$f(x) = a_0 + a_1x^1 + a_2x^2 + \cdots + a_nx^n$$

and

$$g(x) = b_0 + b_1x^1 + b_2x^2 + \cdots + b_nx^n$$

we may define a new polynomial denoted by $(f+g)(x)$, and referred to as the *sum* of $f(x)$ and $g(x)$, by termwise addition of $f(x)$ and $g(x)$ as follows:

$$(f+g)(x) = (b_0 + a_0) + (b_1 + a_1)x^1 + (b_2 + a_2)x^2 + \cdots (b_n + a_n)x^n.$$

Changing the order of summation, we see that

$$(f+g)(x) = \sum_{i=0}^{n}(a_i + b_i)x^i = \sum_{i=0}^{n} a_ix^i + \sum_{i=0}^{n} b_ix^i = f(x) + g(x).$$

We can thus define the polynomial $(f+g)(x)$ being the sum of $f(x)$ and $g(x)$ by the formula

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

We will below extend this definition to general functions.

EXAMPLE 10.13. If $f(x) = 1+x^2-x^4+2x^5$ and $g(x) = 33x+7x^2+2x^5$, then

$$(f+g)(x) = 1 + 33x + 8x^2 - x^4 + 4x^5,$$

where of course we "fill in" the "missing" monomials, i.e. those with coefficients equal to zero in order to use the definition.

In general, to add the polynomials

$$f(x) = \sum_{i=0}^{n} a_ix^i$$

of degree n (assuming that $a_n \neq 0$) and the polynomial

$$g(x) = \sum_{i=0}^{m} b_ix^i$$

of degree m, where we assume that $m \leq n$, we just fill in the "missing" coefficients in g by setting $b_{m+1} = b_{m+2} = \cdots b_n = 0$, and add using the definition.

EXAMPLE 10.14.

$$\sum_{i=0}^{15}(i+1)x^i + \sum_{i=0}^{30} x^i = \sum_{i=0}^{30} a_ix^i$$

with

$$a_i = \begin{cases} i+2, & 0 \leq i \leq 15 \\ i, & 16 \leq i \leq 30 \end{cases}$$

10.6.3 Multiplication of a polynomial by a number

Given a polynomial

$$f(x) = \sum_{i=0}^{n} a_i x^i,$$

and a number $c \in \mathbb{Q}$ we define a new polynomial denoted by $(cf)(x)$, and referred to as the *product* of $f(x)$ *by the number* c, as follows:

$$(cf)(x) = \sum_{i=0}^{n} c a_i x^i.$$

We note that we can equivalently define $(cf)(x)$ by

$$(cf)(x) = cf(x) = c \times f(x).$$

EXAMPLE 10.15.

$$2.3(1 + 6x - x^7) = 2.3 + 13.8x - 2.3x^7.$$

10.6.4 Equality of polynomials

Following these definitions, we say that two polynomials $f(x)$ and $g(x)$ are equal if $(f-g)(x)$ is the zero polynomial with all coefficients equal to zero, that is the coefficients of $f(x)$ and $g(x)$ are the same. Two polynomials are not necessarily equal because they happen to have the same value at just one point!

EXAMPLE 10.16. $f(x) = x^2 - 4$ and $g(x) = 3x - 6$ are both zero at $x = 2$ but are not equal.

10.6.5 Linear combinations of polynomials

We may now combine polynomials by adding them and multiplying them by rational numbers, and thereby obtain new polynomials. Thus, if $f_1(x)$, $f_2(x)$, $\cdots$, $f_n(x)$ are n given polynomials and c_1, $\cdots$, c_n are n given numbers, then

$$f(x) = \sum_{m=1}^{n} c_m f_m(x)$$

is a new polynomial called the *linear combination* of the polynomials f_1, $\cdots$, f_n with *coefficients* c_1, $\cdots$, c_n.

EXAMPLE 10.17. The linear combination of $2x^2$ and $4x - 5$ with coefficients 1 and 2 is

$$1\big(2x^2\big) + 2\big(4x - 5\big) = 2x^2 + 8x - 10.$$

A general polynomial

$$f(x) = \sum_{i=0}^{n} a_i x^i$$

can be described as a linear combination of the particular polynomials 1, x, x^2, $\cdots$, x^n, which are called the *monomials*, see Fig. 10.11 below, with the coefficients a_0, a_1,...,a_n. To make the notation consistent, we set $x^0 = 1$ for all x.

We sum up:

Theorem 10.1 *A linear combination of polynomials is a polynomial. A general polynomial is a linear combination of monomials.*

As a consequence of the definitions made, we get a number of rules for linear combinations of polynomials that reflect the corresponding rules for rational numbers. For example, if f, g and h are polynomials and c is rational number, then

$$f + g = g + f, \tag{10.4}$$
$$(f + g) + h = f + (g + h), \tag{10.5}$$
$$c(f + g) = cf + cg, \tag{10.6}$$

where the variable x was omitted for simplicity.

10.6.6 Multiplication of polynomials

We now go into *multiplication* of polynomials. Given two polynomials $f(x) = \sum_{i=0}^{n} a_i x^i$ and $g(x) = \sum_{j=0}^{m} b_j x^j$, we define a new polynomial denoted by $(fg)(x)$, and referred to as the product of $f(x)$ and $g(x)$, as follows

$$(fg)(x) = f(x)g(x).$$

To see that this is indeed a polynomial we consider first the product of two monomials $f(x) = x^j$ and $g(x) = x^i$:

$$(fg)(x) = f(x)g(x) = x^j x^i = x^j \times x^i = x^{j+i}.$$

We see that the degree of the product is the sum of the degrees of the monomials.

Next, if $f(x) = x^j$ and a polynomial $g(x) = \sum_{i=0}^{n} a_i x^i$, then by distributing x^j, we get

$$\begin{aligned}(fg)(x) = x^j g(x) &= a_0 x^j + a_1 x^j \times x + a_2 x^j \times x^2 + \cdots + a_n x^j \times x^n \\ &= a_0 x^j + a_1 x^{1+j} + a_2 x^{2+j} + \cdots + a_n x^{n+j} \\ &= \sum_{i=0}^{n} a_i x^{i+j},\end{aligned}$$

which is a polynomial of degree $n + j$.

EXAMPLE 10.18.

$$x^3(2 - 3x + x^4 + 19x^8) = 2x^3 - 3x^4 + x^7 + 19x^{11}.$$

Finally, for two general polynomials $f(x) = \sum_{i=0}^{n} a_i x^i$ and $g(x) = \sum_{j=0}^{m} b_j x^j$, we have

$$\begin{aligned}(fg)(x) = f(x)g(x) &= (\sum_{i=0}^{n} a_i x^i)(\sum_{i=0}^{m} b_i x^i) \\ &= \sum_{i=0}^{n} \left(a_i x^i \sum_{j=0}^{m} b_j x^j \right) = \sum_{i=0}^{n} \left(a_i \sum_{j=0}^{m} b_j x^{i+j} \right) \\ &= \sum_{i=0}^{n} \sum_{j=0}^{m} a_i b_j x^{i+j}.\end{aligned}$$

which is a polynomial of degree $n + m$. We consider an example

EXAMPLE 10.19.

$$\begin{aligned}(1 + 2x + 3x^2)(x - x^5) &= 1(x - x^5) + 2x(x - x^5) + 3x^2(x - x^5) \\ &= x - x^5 + 2x^2 - 2x^6 + 3x^3 - 3x^7 \\ &= x + 2x^2 + 3x^3 - x^5 - 2x^6 - 3x^7\end{aligned}$$

We sum up:

Theorem 10.2 *The product of a polynomial of degree n and a polynomial of degree m is a polynomial of degree $n + m$.*

The usual commutative, associative, and distributive laws hold for multiplication of polynomials f, g, and h:

$$fg = gf, \tag{10.7}$$

$$(fg)h = f(gh), \tag{10.8}$$

$$(f + g)h = fh + gh, \tag{10.9}$$

where we again left out the variable x.

Products are tedious to compute but luckily it is not necessary very often and if the polynomials are complicated, we can use *MAPLE*© to compute them for example. There are a couple of examples that are good to keep in mind:

$$\begin{aligned}(x + a)^2 = (x + a)(x + a) &= x^2 + 2ax + a^2 \\ (x + a)(x - a) &= x^2 - a^2 \\ (x + a)^3 &= x^3 + 3ax^2 + 3a^2 x + a^3\end{aligned}$$

10.7 Graphs of general polynomials

A general polynomial of degree greater than 2 or 3 can be a quite complicated function and it is difficult to say much specific about their plots. We show an example in Fig. 10.10. When the degree of a polynomial is large,

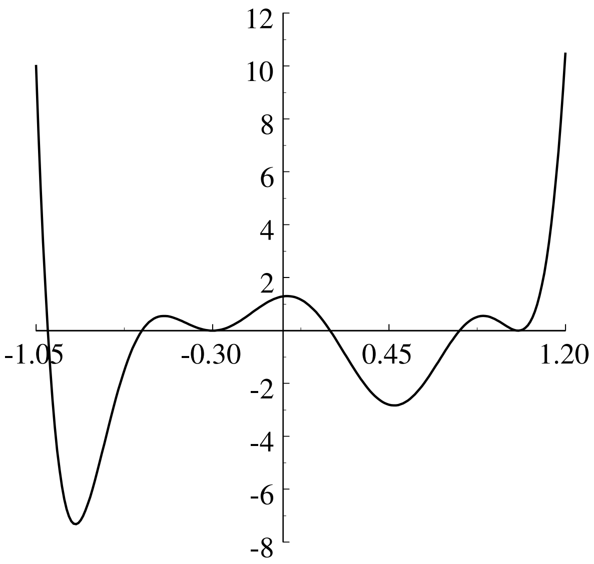


FIGURE 10.10. A plot of $y = 1.296 + 1.296x - 35.496x^2 - 57.384x^3 + 177.457x^4 + 203.889x^5 - 368.554x^6 - 211.266x^7 + 313.197x^8 + 70.965x^9 - 97.9x^{10} - 7.5x^{11} + 10x^{12}$.

the tendency is for the plot to have large "wiggles" which makes it difficult to plot the function. The value of the polynomial shown in Fig. 10.10 is 987940.8 at $x = 3$.

On the other hand, we can plot the monomials rather easily. It turns out that once the degree $n \geq 2$, the plots of the monomials with even degree n all have a similar shape, as do the plots of all the monomials with odd degree. We show some samples in Fig. 10.11. One of the most obvious feature of the graphs of the monomials are the symmetry in the plots. When the degree is even, the plots are symmetric across the y-axis, see Fig. 10.12. This means that the value of the monomial is the same for x and $-x$, or in other words $x^m = (-x)^m$ for m even. When the degree is odd, the plots are symmetric through the origin. In other words, the value of the function for x is the negative of the value of the function for $-x$ or $(-x)^m = -x^m$ for m odd.

We can use the ideas of scaling and translation to graph functions of the form $y = a(x - x_0)^m + d$.

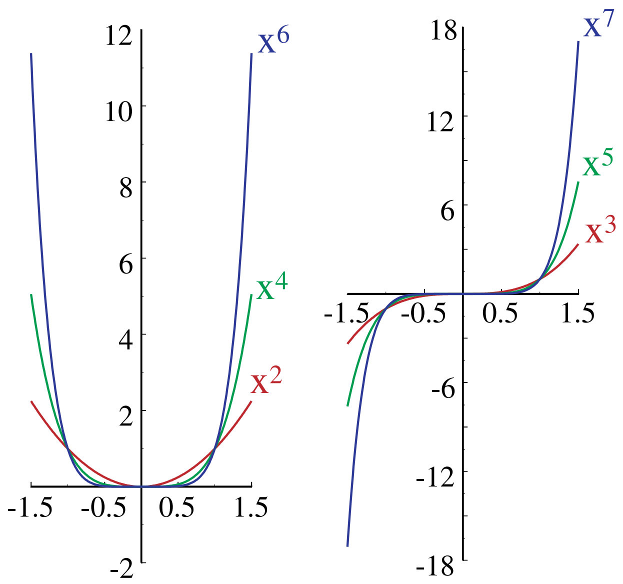


FIGURE 10.11. Plots of some monomials.

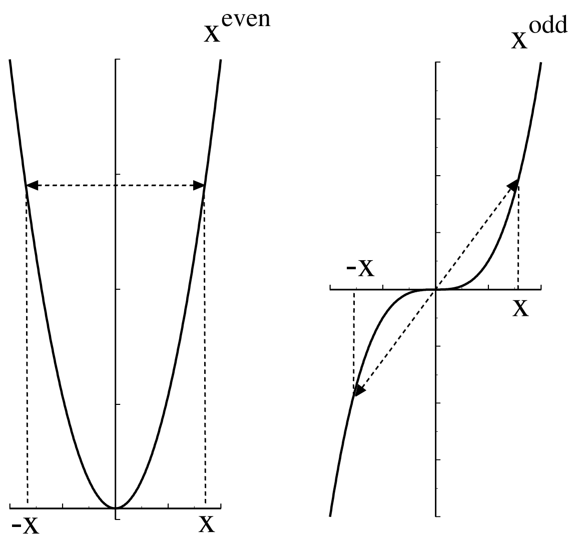


FIGURE 10.12. The symmetries of the monomial functions of even and odd degree.

EXAMPLE 10.20. We plot $y = -.5(x-1)^3 - 6$ in Fig. 10.13 by systematically using translations and scaling.

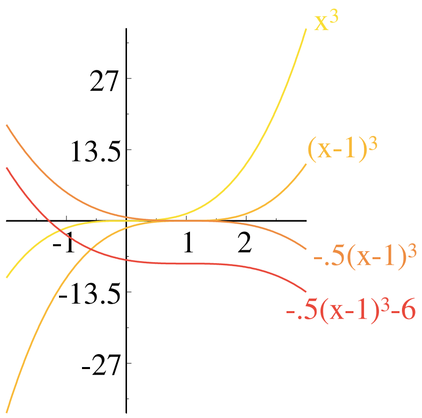


FIGURE 10.13. The procedure for plotting $y = -.5(x-1)^3 - 6$.

Luckily, however, there is no procedure like completing the square for monomials of higher degree.

10.8 Piecewise polynomial functions

We started this chapter by declaring that polynomials are building blocks for the mathematics of functions. An important class of functions constructed using polynomials are the *piecewise polynomials*. These are functions that are equal to polynomials on intervals contained in the domain.

We have already met one example, namely

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

The function $|x|$ looks like $y = x$ for $x \geq 0$ and $y = -x$ for $x < 0$. We plot it in Fig. 10.14. The most interesting thing to note about the graph of $|x|$ is the sharp corner at $x = 0$, which occurs right at the transition point of this piecewise polynomial.

Chapter 10 Problems

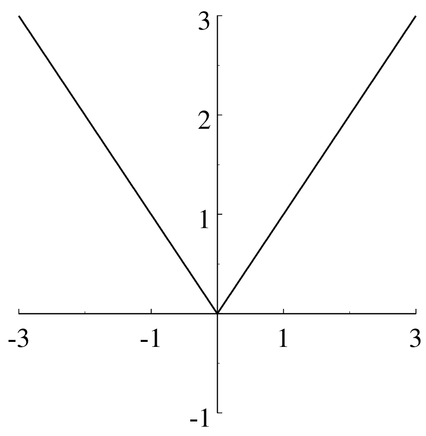


FIGURE 10.14. Plot of $y = |x|$.

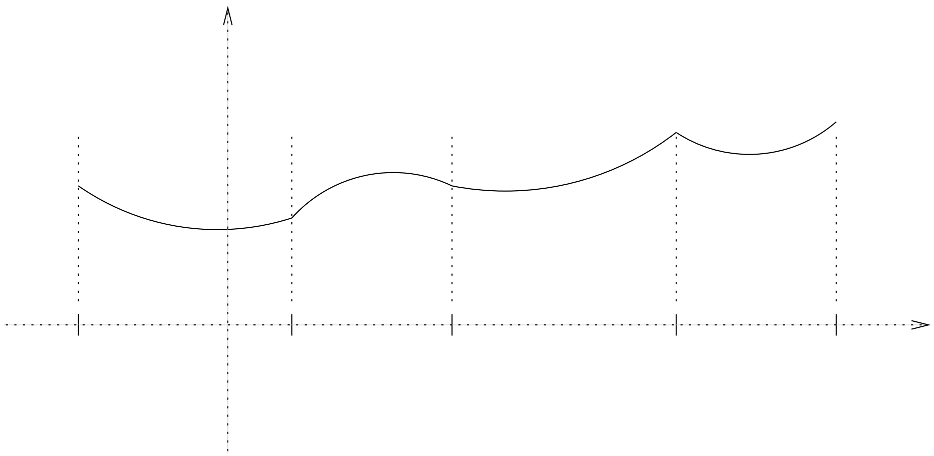


FIGURE 10.15. Plot of a piecewise (quadratic) polynomial function.

10.1. Find the point-slope equations of the lines passing through the following pairs of points. Plot the line in each case.

(a) $(1, 3)$ & $(2, 7)$ (b) $(-4, 2)$ & $(-6, 3)$

(c) $(3, 7)$ & $(5, 7)$ (d) $(3.5, 1.5)$ & $(2.1, 11.8)$

(e) $(-3, 2)$ & $(-3, 3)$ (f) $(2, -1)$ & $(4, -7)$.

10.2. Find the slope-intercept equations of the lines passing through the following pairs of points. Plot the line in each case.

(a) $(4, -6)$ & $(14, 2)$ (b) $(3, -2)$ & $(-1, 4)$

(c) $(13, 4)$ & $(13, 89)$ (d) $(4, 4)$ & $(6, 4)$

(e) $(-.2, 9)$ & $(-.4, 7)$ (f) $(-1, -1)$ & $(-4, 7)$.

10.3. Find a formula for the x-intercept of a line given in the form $y = mx + b$ in terms of m and b.

10.4. Plot the lines $y = \frac{1}{2}x$, $y = \frac{1}{2}x - 2$, $y = \frac{1}{2}x + 4$, $y = \frac{1}{2}x + 1$ using translation.

10.5. Are the lines $2 - y = 7(4 - x)$ and $y = 7x - 13$ parallel?

10.6. Are the lines $y = \frac{3}{11}x - 4$ and $y = 13 - \frac{11}{3}x$ perpendicular?

10.7. Find the point of intersection of the following pairs of lines:
(a) $y = 3x + 2$ and $y = -4x - 2$,
(b) $y - 5 = 7(x - 1)$ and $y + 3 = -4(x - 9)$.

10.8. Find the lines that are (a) parallel and (b) perpendicular to the line through $(9, 4)$ and $(-1, 3)$ and passing through the point $(3, 0)$.

10.9. Find the lines that are (a) parallel and (b) perpendicular to the line through $(-2, 7)$ and $(8, 8)$ and passing through the point $(1, 2)$.

10.10. Show that $f(x) = x^2$ is decreasing for $x < 0$.

10.11. Plot the following quadratic functions for $-2 \le x \le 2$:

(a) $6x^2$ (b) $-\frac{1}{4}x^2$ (c) $\frac{4}{3}x^2$.

10.12. Plot the following quadratic functions for $-3 \le x \le 3$:

(a) $(x - 2)^2$ (b) $(x + 1.5)^2$ (c) $(x + .5)^2$.

10.13. Plot the following quadratic functions for $-2 \le x \le 2$:

(a) $x^2 - 3$ (b) $x^2 + 2$ (c) $x^2 - .5$

10.14. Plot the following quadratic functions for $-3 \le x \le 3$:

(a) $-\frac{1}{2}(x-1)^2 + 2$ (b) $2(x+2)^2 - 5$ (c) $\frac{1}{3}(x-3)^2 - 1$.

10.15. Complete the square on the following quadratic functions then plot them for $-3 \le x \le 3$:

(a) $x^2 + 4x + 5$ (b) $2x^2 - 2x - \frac{1}{2}$ (c) $-\frac{1}{3}x^2 + 2x - 1$.

10.16. Write the following finite sums using the summation notation. Be sure to get the starting and ending values for the index correct!

(a) $1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots + \frac{1}{n^2}$ (b) $-1 + \frac{1}{4} - \frac{1}{9} + \frac{1}{16} - \cdots \pm \frac{1}{n^2}$

(c) $1 + \frac{1}{2\times 3} + \frac{1}{3\times 4} + \cdots + \frac{1}{n\times(n+1)}$ (d) $1 + 3 + 5 + 7 + \cdots + 2n + 1$

(e) $x^4 + x^5 + \cdots + x^n$ (f) $1 + x^2 + x^4 + x^6 + \cdots + x^{2n}$.

10.17. Write the finite sum $\sum_{i=1}^{n} i^2$ so that: (a) i starts with -1, (b) i starts with 15, (c) the coefficent has the form $(i+4)^2$, (d) i ends with $n+7$.

10.18. Given $f_1(x) = -4 + 6x + 7x^3$, $f_2(x) = 2x^2 - x^3 + 4x^5$ and $f_3(x) = 2 - x^4$, compute the following polynomials

(a) $f_1 - 4f_2$ (b) $3f_2 - 12f_1$ (c) $f_2 + f_1 + f_3$

(d) $f_2 f_1$ (e) $f_1 f_3$ (f) $f_2 f_3$.

(g) $f_1 f_3 - f_2$ (h) $(f_1 + f_2) f_3$ (i) $f_1 f_2 f_3$

10.19. For a equal to a constant, compute

(a) $(x+a)^2$ (b) $(x+a)^3$ (c) $(x-a)^3$ (d) $(x+a)^4$.

10.20. Compute $f_1 f_2$ where $f_1(x) = \sum_{i=0}^{8} i^2 x^i$ and $f_2(x) = \sum_{j=0}^{11} \frac{1}{j+1} x^j$.

10.21. Plot the function

$$f(x) = 360x - 942x^2 + 949x^3 - 480x^4 + 130x^5 - 18x^6 + x^7$$

using Matlab or Maple. This takes some trial and error in choosing a good interval on which to plot. You should make plots on several different intervals, starting with $-.5 \le x \le .5$ then increasing the size.

10.22. (a) Show that the monomial x^3 is increasing for all x. (b) Show the monomial x^4 is decreasing for $x < 0$ and increasing for $x > 0$.

10.23. Plot the following monomial functions for $-3 \leq x \leq 3$:

$$\text{(a) } x^3 \qquad \text{(b) } x^4 \qquad \text{(c) } x^5.$$

10.24. Plot the following polynomials for $-3 \leq x \leq 3$:

$$\text{(a) } \tfrac{1}{3}(x+2)^3 - 2 \qquad \text{(b) } 2(x-1)^4 - 13 \qquad \text{(c) } (x+1)^5 - 1.$$

10.25. Plot the following piecewise polynomials for $-2 \leq x \leq 2$

$$\text{(a) } f(x) = \begin{cases} 1, & -2 \leq x \leq -1, \\ x^2, & -1 < x < 1, \\ x, & 1 \leq x \leq 2. \end{cases} \quad \text{(b) } f(x) = \begin{cases} -1-x, & -2 \leq x \leq -1, \\ 1+x, & -1 < x \leq 0, \\ 1-x, & 0 < x \leq 1, \\ -1+x, & 1 < x \leq 2. \end{cases}$$

11
Combinations of functions

> And he gave a deep sigh, and tried very hard to listen to what Owl was saying. (Winnie-the Pooh)

11.1 Introduction

In this chapter we consider different ways of creating new functions by combining old ones. We often seek to describe complicated functions as combinations of simpler functions that we know. In the last chapter, we saw how a general polynomials can be created adding up multiples of monomials, that is, as linear combinations of monomials. In this chapter, we consider first linear combinations of arbitrary functions, then multiplication and division, and finally composition of functions.

The idea of combining simple things to get complex ones is fundamental in many different settings. Music is a good example: chords or harmonies are formed by combining single tones, complex rhythmic patterns may be formed by overlaying simple basic rhythmic patterns, single instruments are combined to form an orchestra. Another example is a fancy dinner that is made up of an entree, main dish, dessert, coffee, together with aperitif, wines and cognac, in endless combinations. Moreover, each dish is formed by combining ingredients like beef, carrots and potatoes.

11.2 Sum of two functions and product of a function with a number

Given two functions $f_1 : D(f_1) \to \mathbb{Q}$ and $f_2 : D(f_2) \to \mathbb{Q}$, we define a new function denoted by $(f_1 + f_2)(x)$, and referred to as the *sum* of $f_1(x)$ and $f_2(x)$, as follows

$$(f_1 + f_2)(x) = f_1(x) + f_2(x), \quad \text{for } x \in D(f_1) \cap D(f_2).$$

Of course, we have to assume that x belongs to both $D(f_1)$ and $D(f_2)$ for both $f_1(x)$ and $f_2(x)$ to be defined. We can thus write $D(f_1 + f_2) = D(f_1) \cap D(f_2)$.

Further, given a function $f : D(f) \to \mathbb{Q}$ and a number $c \in \mathbb{Q}$, we define a new function denoted by $(cf)(x)$, and referred to as the *product* of $f(x)$ with c, as follows

$$(cf)(x) = cf(x) \quad \text{for } x \in D(f).$$

The domain of cf is equal to the domain of f, that is, $D(cf) = D(f)$.

The definitions of sum of functions and product of a function with a number are consistent with the corresponding definitions for polynomials made above.

EXAMPLE 11.1. The function $f(x) = x^3 + 1/x$ defined on $D(f) = \{x \in \mathbb{Q} : x \neq 0\}$ is the sum of the functions $f_1(x) = x^3$ with domain $D(f_1) = \mathbb{Q}$ and $f_2(x) = 1/x$ with domain $D(f_2) = \{x \text{ in } \mathbb{Q} : x \neq 0\}$. The function $f(x) = x^2 + 2^x$ defined on $\mathbb{Z}$ is the sum of x^2 defined on $\mathbb{Q}$ and 2^x defined on $\mathbb{Z}$.

11.3 Linear combinations of functions

Given n functions $f_1 : D(f_1) \to \mathbb{Q}$, ..., $f_n : D(f_n) \to \mathbb{Q}$, and numbers c_1,...,c_n, we define the *linear combination* of $f_1, ..., f_n$ with coefficients $c_1, ..., c_n$, denoted by $(c_1 f + \cdots + c_n f_n)(x)$, as follows

$$(c_1 f + \cdots + c_n f_n)(x) = c_1 f_1(x) + \cdots + c_n f_n(x)$$

The domain $D(c_1 f + \cdots + c_n f_n)$ of the linear combination $c_1 f + \cdots + c_n f_n$ is the intersection of the domains $D(f_1), \cdots, D(f_n)$.

EXAMPLE 11.2. The domain of the linear combination of $\left\{\frac{1}{x}, \frac{x}{1+x}, \frac{1+x}{2+x}\right\}$ given by

$$-\frac{1}{x} + 2\frac{x}{1+x} + 6\frac{1+x}{2+x}$$

is $\{x \text{ in } \mathbb{Q} : x \neq 0, x \neq -1, x \neq -2\}$.

The sigma notation is useful for writing general linear combinations.

EXAMPLE 11.3. The linear combination of $\{\frac{1}{x}, \cdots, \frac{1}{x^n}\}$ given by

$$\frac{2}{x}+\frac{4}{x}+\frac{8}{x}+\cdots+\frac{2^n}{x^n}=\sum_{i=1}^{n}\frac{2^i}{x^i}$$

has domain $\{x \text{ in } \mathbb{Q} : x \neq 0\}$.

11.4 Multiplication and division of functions

We multiply functions using the same idea used to multiply polynomials. Given two functions $f_1 : D(f_1) \to \mathbb{Q}$ and $f_2 : D(f_2) \to \mathbb{Q}$ we define the *product* function $(f_1 f_2)(x)$ by

$$(f_1 f_2)(x) = f_1(x) f_2(x) \quad \text{for } x \in D(f_1) \cap D(f_2),$$

and *quotient* function by

$$(f_1/f_2)(x) = \frac{f_1}{f_2}(x) = \frac{f_1(x)}{f_2(x)} \quad \text{for } x \in D(f_1) \cap D(f_2),$$

where we of course also assume that $f_2(x) \neq 0$.

EXAMPLE 11.4. The function

$$f(x) = (x^2-3)^3\big(x^6 - \frac{1}{x} - 3\big)$$

with $D(f) = \{x \in \mathbb{Q} : x \neq 0\}$ is the product of the functions $f_1(x) = (x^2-3)^3$ and $f_2(x) = x^6 - 1/x - 3$. The function $f(x) = x^2\, 2^x$ is the product of x^2 and 2^x.

EXAMPLE 11.5. The domain of

$$\frac{1+1/(x+3)}{2x-5}$$

is the intersection of $\{x \text{ in } \mathbb{Q} : x \neq -3\}$ and $\{x \text{ in } \mathbb{Q}\}$ excepting $x = 5/2$ or $\{x \text{ in } \mathbb{Q} : x \neq -3, 5/2\}$.

11.5 Rational functions

The quotient f_1/f_2 of two polynomials $f_1(x)$ and $f_2(x)$ is called a *rational function.* This is the analog of a rational number which is the quotient of two integers.

Example 11.6. The function $f(x) = 1/x$ is a rational function defined for $\{x \text{ in } \mathbb{Q} : x \neq 0\}$. The function

$$f(x) = \frac{(x^3 - 6x + 1)(x^{11} - 5x^6)}{(x^4 - 1)(x + 2)(x - 5)}$$

is a rational function defined on $\{x \text{ in } \mathbb{Q} : x \neq 1, -1, -2, 5\}$.

In an example above, we saw that $x - 3$ divides into $x^2 - 2x - 3$ exactly because $x^2 - 2x - 3 = (x - 3)(x + 1)$ so

$$\frac{x^2 - 2x - 3}{x - 3} = x + 1.$$

In the same way, a rational number p/q sometimes simplifies to an integer, in other words q divides into p exactly without a remainder. We can determine if this is true by using long division. It turns out that long division also works for polynomials. Recall that in long division, we match the leading digit of the denominator with the remainder at each stage. When dividing polynomials, we write them as a linear combination of monomials starting with the monomial of highest degree and then match coefficients of the monomials one by one.

Example 11.7. We show a couple of examples of polynomial division. In Fig. 11.1, we give an example where the remainder is zero. We

$$\begin{array}{r l}
 & x^2 + 5x + 3 \\ \cline{2-2}
x-1 \big) & x^3 + 4x^2 - 2x - 3 \\
 & \underline{x^3 - x^2} \\
 & \quad 5x^2 - 2x \\
 & \quad \underline{5x^2 - 5x} \\
 & \qquad 3x^2 - 3x \\
 & \qquad \underline{3x^2 - 3x} \\
 & \qquad\qquad 0
\end{array}$$

FIGURE 11.1. An example of polynomial division with no remainder.

conclude that

$$\frac{x^3 + 4x^2 - 2x + 3}{x - 1} = x^2 + 5x + 3.$$

In Fig. 11.2, we give an example in which there is a non-zero remainder, i.e. we carry out the division to the point where the remaining numerator has lower degree than the denominator. Note that in this example,

$$
\begin{array}{r|l}
 & \quad 2x^2 - 2x + 15 \\
\cline{2-2}
x^2+x-3 & 2x^4 + 0x^3 + 7x^2 - 8x + 3 \\
 & 2x^4 + 2x^3 - 6x^2 \\
\cline{2-2}
 & \quad - 2x^3 + 13x^2 - 8x \\
 & \quad - 2x^3 - 2x^2 + 6x \\
\cline{2-2}
 & \qquad 15x^2 - 14x + 3 \\
 & \qquad 15x^2 + 15x - 45 \\
\cline{2-2}
 & \qquad\qquad -29x + 48
\end{array}
$$

FIGURE 11.2. An example of polynomial division with a remainder.

the numerator is "missing" a term so we fill in the missing term with a zero coefficient to make the division easier. We conclude that

$$\frac{2x^4 + 7x^2 - 8x + 3}{x^2 + x - 3} = 2x^2 - 2x + 15 + \frac{-29x + 48}{x^2 + x - 3}.$$

We shall now consider polynomial division in the special case of a denominator of the form $x - \bar{x}$ of degree one, where $\bar{x}$ is considered fixed, resulting in

$$f(x) = (x - \bar{x})\, g(x) + r(x), \tag{11.1}$$

where the reminder polynomial $r(x)$ now must be of degree zero, that is a constant.

The following result is of particular interest. If $f(x)$ is a polynomial of degree n with $f(\bar{x}) = 0$, then $x - \bar{x}$ is a *factor* of $f(x)$, that is, division of $f(x)$ with $x - \bar{x}$ gives

$$f(x) = (x - \bar{x})\, g(x) + r(x) \tag{11.2}$$

with $r(x) \equiv 0$. Conversely, if $r(x) \equiv 0$, then obviously $f(\bar{x}) = 0$. For the proof of this we note that the degree of $r(x)$ is less than the degree of $x - \bar{x}$, that is $r(x)$ is in fact a constant. Further $r(\bar{x}) = 0$ because $f(\bar{x}) = 0$. That is $r(x)$ is a constant which is zero, that is $r(x) \equiv 0$. We have thus proved

Theorem 11.1 *If $\bar{x}$ is a root of a polynomial $f(x)$, that is if $f(\bar{x}) = 0$, then $f(x)$ factors as $f(x) = (x - \bar{x})g(x)$ for some polynomial $g(x)$ of degree one less than the degree of $f(x)$. The factor $g(x)$ can be found by polynomial division of $f(x)$ by $x - \bar{x}$.*

11.6 The composition of functions

Given two functions f_1 and f_2, we can define a new function f by first applying f_1 to an input and then applying f_2 to the result, i.e.

$$f(x) = f_2(f_1(x))$$

We say that f is the *composition* of f_2 and f_1 and we write $f = f_2 \circ f_1$, that is

$$(f_2 \circ f_1)(x) = f_2(f_1(x)).$$

We illustrate this operation in Fig. 11.3.

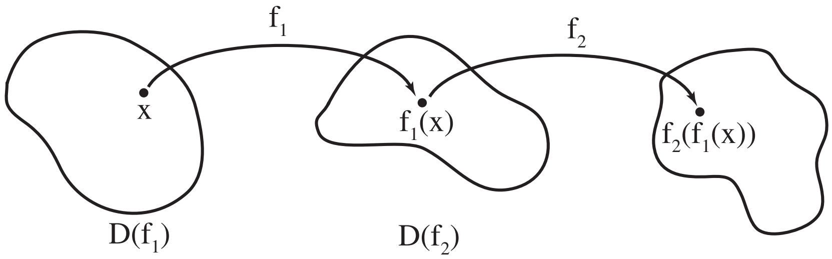


FIGURE 11.3. Illustration of the composition $f_2 \circ f_1$.

EXAMPLE 11.8. If $f_1(x) = x^2$ and $f_2(x) = x + 1$ then $f_1 \circ f_2(x) = f_1(f_2(x)) = (x+1)^2$ while $f_2 \circ f_1 = f_2(f_1(x)) = x^2 + 1$.

This example illustrates the general fact that $f_2 \circ f_1 \neq f_1 \circ f_2$ in most cases.

Determining the domain of the composition of $f_2 \circ f_1$ can be complicated. Certainly to compute $f_2(f_1(x))$, we have to make certain that x is in the domain of f_1 otherwise $f_1(x)$ will be undefined. Next we apply f_2 to the result, therefore $f_1(x)$ must have a value that is in the domain of f_2. Therefore the domain of $f_2 \circ f_1$ is the set of points x in $D(f_1)$ such that $f_1(x)$ is in $D(f_2)$.

EXAMPLE 11.9. Let $f_1(x) = 3 + 1/x^2$ and $f_2(x) = 1/(x-4)$. Then $D(f_1) = \{x \text{ in } \mathbb{Q} : x \neq 0\}$ while $D(f_2) = \{x \text{ in } \mathbb{Q} : x \neq 4\}$. Therefore to compute $f_2 \circ f_1$, we must avoid any points where $3 + 1/x^2 = 4$ or $1/x^2 = 1$ or $x = 1$ and $x = -1$. We conclude that $D(f_2 \circ f_1) = \{x \text{ in } \mathbb{Q} : x \neq 0, 1, -1\}$.

Chapter 11 Problems

11.1. Determine the domains of the following functions

(a) $3(x-4)^3+2x^2+\dfrac{4x}{3x-1}+\dfrac{6}{(x-1)^2}$ (d) $\dfrac{(2x-3)\frac{2}{x}}{4x+6}$

(b) $2+\dfrac{4}{x}-\dfrac{6x+4}{(x-2)(2x+1)}$ (e) $\dfrac{6x-1}{(2-3x)(4+x)}$

(c) $x^3\left(1+\dfrac{1}{x}\right)$ (f) $\dfrac{4}{x+2}+\dfrac{6}{x^2+3x+2}$

11.2. Write the following linear combinations using the sigma notation and determine the domain of the result.

(a) $2x(x-1)+3x^2(x-1)^2+4x^3(x-1)^3+\cdots+100x^{101}(x-1)^{101}$

(b) $\dfrac{2}{x-1}+\dfrac{4}{x-2}+\dfrac{8}{x-3}+\cdots+\dfrac{8192}{x-13}$

11.3. (a) Let $f(x)=ax+b$, where a and b are numbers and show that $f(x+y)=f(x)+f(y)$ for *all* numbers x and y. (b) Let $g(x)=x^2$ and show that $g(x+y)\neq g(x)+g(y)$ unless x and y have special values.

11.4. Use polynomial division on the following rational functions to show that the denominator divides the numerate exactly or to compute the remainder if not.

(a) $\dfrac{x^2+2x-3}{x-1}$ (b) $\dfrac{2x^2-7x-4}{2x+1}$

(c) $\dfrac{4x^2+2x-1}{x+6}$ (d) $\dfrac{x^3+3x^2+3x+2}{x+2}$

(e) $\dfrac{5x^3+6x^2-4}{2x^2+4x+1}$ (f) $\dfrac{x^4-4x^2-5x-4}{x^2+x+1}$

(g) $\dfrac{x^8-1}{x^3-1}$ (h) $\dfrac{x^n-1}{x-1}$, n in $\mathbb{N}$

11.5. Given $f_1(x)=3x-5$, $f_2(x)=2x^2+1$, and $f_3(x)=4/x$, write out formulas for the following functions

(a) $f_1\circ f_2$ (b) $f_2\circ f_3$ (c) $f_3\circ f_1$ (d) $f_1\circ f_2\circ f_3$

11.6. With $f_1(x)=4x+2$ and $f_2(x)=x/x^2$, show that $f_1\circ f_2\neq f_2\circ f_1$.

11.7. Let $f_1(x)=ax+b$ and $f_2(x)=cx+d$ where a, b, c, and d are rational numbers. Find a condition on the numbers a, b, c, and d that implies that $f_1\circ f_2=f_2\circ f_1$ and produce an example that satisfies the condition.

11.8. For the given functions f_1 and f_2, determine the domain of $f_2\circ f_1$

(a) $f_1(x)=4-\dfrac{1}{x}$ and $f_2(x)=\dfrac{1}{x^2}$

(b) $f_1(x)=\dfrac{1}{(x-1)^2}-4$ and $f_2(x)=\dfrac{x+1}{x}$

12
Lipschitz continuity

> Calculus required continuity, and continuity was supposed to require the infinitely little, but nobody could discover what the infinitely little might be. (Russell)

12.1 Introduction

When we graph a function $f(x)$ of a rational variable x, we make a leap of faith and assume that the function values $f(x)$ vary "smoothly" or "continuously" between the sample points x, so that we can draw the graph of the function without lifting the pen. In particular, we assume that the function value $f(x)$ does not make unknown sudden jumps for some values of x. We thus assume that the function value $f(x)$ changes by a small amount if we change x by a small amount. A basic problem in Calculus is to measure how much the function values $f(x)$ may change when x changes, that is, to measure the "degree of continuity" of a function. In this chapter, we approach this basic problem using the concept of *Lipschitz continuity*, which plays a basic role in the version of Calculus presented in this book.

There will be a lot of inequalities ($<$ and $\leq$) and absolute values ($|\cdot|$) in this chapter, so it might be a good idea before you start to review the rules for operating with these symbols from Chapter *Rational numbers*.

FIGURE 12.1. Rudolph Lipschitz (1832-1903), Inventor of Lipschitz continuity: "Indeed, I have found a very nice way of expressing continuity....".

12.2 The Lipschitz continuity of a linear function

To start with we consider the behavior of a linear polynomial. The value of a constant polynomial doesn't change when we change the input, so the linear polynomial is the first interesting example to consider. Suppose the linear function is $f(x) = mx + b$, with $m \in \mathbb{Q}$ and $b \in \mathbb{Q}$ given, and let $f(x_1) = mx_1 + b$ and $f(x_2) = mx_2 + b$ to be the function values values for $x = x_1$ and $x = x_2$. The change in the input is $|x_2 - x_1|$ and for the corresponding change in the output $|f(x_1) - f(x_2)|$, we have

$$|f(x_2) - f(x_1)| = |(mx_2 + b) - (mx_1 + b)| = |m(x_2 - x_1)| = |m||x_2 - x_1|. \tag{12.1}$$

In other words, the absolute value of the change in the function values $|f(x_2) - f(x_1)|$ is proportional to the absolute value of the change in the input values $|x_2 - x_1|$ with constant of proportionality equal to the slope $|m|$. In particular, this means that we can make the change in the output arbitrarily small by making the change in the input small, which certainly fits our intuition that a linear function varies continuously.

EXAMPLE 12.1. Let $f(x) = 2x$ give the total number of miles for an "out and back" bicycle ride that is x miles one way. To increase a given ride by a total of 4 miles, we increase the one way distance x by $4/2 = 2$ miles while to increase a ride by a total of .01 miles, we increase the one way distance x by .005 miles.

We now make an important observation: the slope m of the linear function $f(x) = mx + b$ determines how much the function values change as the input value x changes. The larger $|m|$ is, the steeper the line is, and the more the function changes for a given change in input. We illustrate in Fig. 12.2.

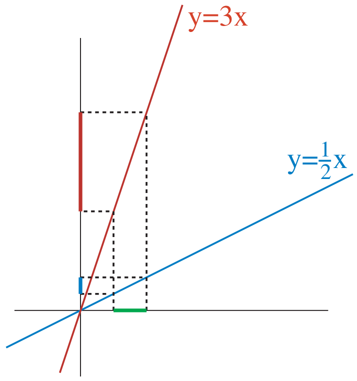


FIGURE 12.2. These two linear functions which change a different amount for a given change in input.

EXAMPLE 12.2. Suppose that $f_1(x) = 4x + 1$ while $f_2(x) = 100x - 5$. To increase the value of $f_1(x)$ at x by an amount of .01, we change the value of x by $.01/4 = .0025$. On the other hand, to change the value of $f_2(x)$ at x by an amount of .01, we change the value of x by $.01/100 = .0001$.

12.3 The definition of Lipschitz continuity

We are now prepared to introduce the concept of Lipschitz continuity, designed to measure change of function values versus change in the independent variable for a general function $f : I \to \mathbb{Q}$ where I is a set of rational numbers. Typically, I may be an interval of rational numbers $\{x \in \mathbb{Q} : a \leq x \leq b\}$ for some rational numbers a and b. If x_1 and x_2 are two numbers in I, then $|x_2 - x_1|$ is the change in the input and $|f(x_2) - f(x_1)|$ is the corresponding change in the output. We say that f is *Lipschitz continuous* with *Lipschitz constant* L_f on I, if there is a (necessarily nonnegative) constant L_f such that

$$|f(x_1) - f(x_2)| \leq L_f |x_1 - x_2| \quad \text{for all } x_1, x_2 \in I. \tag{12.2}$$

As indicated by the notation, the Lipschitz constant L_f depends on the function f, and thus may vary from being small for one function to be large for another function. If L_f is small, then $f(x)$ may change only a little with a small change of x, while if L_f is large, then $f(x)$ may change a lot under only a small change of x. Again: L_f may vary from small to large depending on the function f.

EXAMPLE 12.3. A linear function $f(x) = mx+b$ is Lipschitz continuous with Lipschitz constant $L_f = |m|$ on the entire set of rational numbers $\mathbb{Q}$.

EXAMPLE 12.4. We show that $f(x) = x^2$ is Lipschitz continuous on the interval $I = [-2, 2]$ with Lipschitz constant $L_f = 4$. We choose two rational numbers x_1 and x_2 in $[-2, 2]$. The corresponding change in the function values is

$$|f(x_2) - f(x_1)| = |x_2^2 - x_1^2|.$$

The goal is to estimate this in terms of the difference in the input values $|x_2 - x_1|$. Using the identity for products of polynomials derived in Section 10.6, we get

$$|f(x_2) - f(x_1)| = |x_2 + x_1|\,|x_2 - x_1|. \tag{12.3}$$

We have the desired difference on the right, but it is multiplied by a factor that depends on x_1 and x_2. In contrast, the analogous relationship (12.1) for the linear function has a factor that is constant, namely $|m|$. At this point, we have to use the fact that x_1 and x_2 are in the interval $[-2, 2]$, which means that

$$|x_2 + x_1| \leq |x_2| + |x_1| \leq 2 + 2 = 4,$$

by the triangle inequality. We conclude that

$$|f(x_2) - f(x_1)| \leq 4|x_2 - x_1|$$

for all x_1 and x_2 in $[-2, 2]$.

Lipschitz continuity quantifies the idea of continuous behavior of a function $f(x)$ using the Lipschitz constant L_f. We repeat: If L_f is moderately sized then small changes in input x yield small changes in the function's output $f(x)$, but a large Lipschitz constant means that the function's values $f(x)$ may make a large change when the input values x change by only a small amount.

However it is important to note that there is a certain amount of imprecision inherit to the definition of Lipschitz continuity (12.2) and we have to be circumspect about drawing conclusions when the Lipschitz constant is large. The reason is that (12.2) is only an **upper estimate** on how much the function changes and the actual change might be much smaller than indicated by the constant.

EXAMPLE 12.5. From Example 12.4, we know that $f(x) = x^2$ is Lipschitz continuous on $I = [-2, 2]$ with Lipschitz constant $L_f = 4$. It is also Lipschitz constant on I with Lipschitz constant $L_f = 121$ since

$$|f(x_2) - f(x_1)| \le 4|x_2 - x_1| \le 121|x_2 - x_1|.$$

But the second value of L_f greatly overestimates the change in f, whereas the value $L_f = 4$ is just about right when x_1 and x_2 are near 2 since $2^2 - 1.9^2 = .39 = 3.9 \times (2 - 1.9)$ and $3.9 \approx 4$.

To determine the Lipschitz constant, we have to make some estimates and the result can vary greatly depending on how difficult the estimates are to compute and our skill at making estimates.

It is also important to note that the size and location of the interval in the definition is important and if we change the interval then we expect to get a different Lipschitz constant L_f.

EXAMPLE 12.6. We show that $f(x) = x^2$ is Lipschitz continuous on the interval $I = [2, 4]$, with Lipschitz constant $L_f = 8$. Starting with (12.3), for x_1 and x_2 in $[2, 4]$ we have

$$|x_2 + x_1| \le |x_2| + |x_1| \le 4 + 4 = 8$$

so

$$|f(x_2) - f(x_1)| \le 8|x_2 - x_1|$$

for all x_1 and x_2 in $[2, 4]$.

The reason that the Lipschitz constant is bigger in the second example is clear from the graph, see Fig. 12.3, where we show the change in f corresponding to equal changes in x near $x = 2$ and $x = 4$. Because $f(x) = x^2$ is steeper near $x = 4$, f changes more near $x = 4$ for a given change in input.

EXAMPLE 12.7. $f(x) = x^2$ is Lipschitz continuous on $I = [-8, 8]$ with Lipschitz constant $L_f = 16$ and on $I = [-400, 200]$ with $L_f = 800$.

In all of the examples involving $f(x) = x^2$, we use the fact that the interval under consideration is of finite size. A set of rational numbers I is *bounded* with size a if $|x| \le a$ for all x in I, for some (finite) rational number a.

EXAMPLE 12.8. The set of rational numbers $I = [-1, 500]$ is bounded but the set of even integers is not bounded.

While linear functions are Lipschitz continuous on the unbounded set $\mathbb{Q}$, functions that are not linear are usually only Lipschitz continuous on bounded sets.

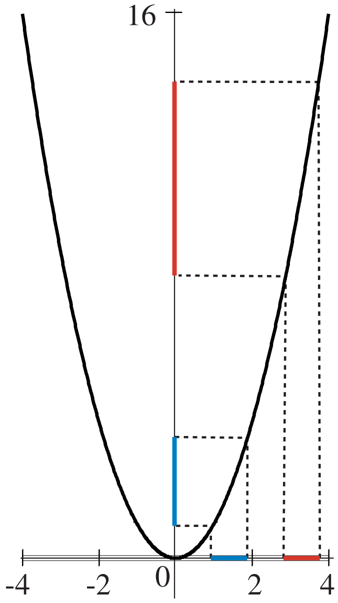


FIGURE 12.3. The change in $f(x) = x^2$ for equal changes in x near $x = 2$ and $x = 4$.

EXAMPLE 12.9. The function $f(x) = x^2$ is **not** Lipschitz continuous on the set $\mathbb{Q}$ of rational numbers. This follows from (12.3) because $|x_1 + x_2|$ can be made arbitrarily large by choosing x_1 and x_2 freely in $\mathbb{Q}$, so it is not possible to find a constant L_f such that

$$|f(x_2) - f(x_1)| = |x_2 + x_1||x_2 - x_1| \le L_f|x_2 - x_1|$$

for all x_1 and x_2 in $\mathbb{Q}$.

The definition of Lipschitz continuity is due to the German mathematician Rudolph Lipschitz (1832-1903), who used his concept of continuity to prove existence of solutions to some important differential equations. This is not the usual definition of continuity used in Calculus courses, which is purely qualitative, while Lipschitz continuity is quantitative. Of course there is a strong connection, and a function which is Lipschitz continuous is also continuous according to the usual definition of continuity, while the opposite may not be true: Lipschitz continuity is a somewhat more demanding property. However, quantifying continuous behavior in terms of Lipschitz continuity simplifies many aspects of mathematical analysis and the use of Lipschitz continuity has become ubiquitous in engineering and applied mathematics. It also has the benefit of eliminating some rather technical issues in defining continuity that are tricky yet unimportant in practice.

12.4 Monomials

Continuing the investigation of continuous functions, we next show that the monomials are Lipschitz continuous on bounded intervals, as we expect based on their graphs.

EXAMPLE 12.10. We show that the function $f(x) = x^4$ is Lipschitz continuous on $I = [-2, 2]$ with Lipschitz constant $L_f = 32$. We choose x_1 and x_2 in I and we want to estimate

$$|f(x_2) - f(x_1)| = |x_2^4 - x_1^4|$$

in terms of $|x_2 - x_1|$.

To do this we first show that

$$x_2^4 - x_1^4 = (x_2 - x_1)(x_2^3 + x_2^2 x_1 + x_2 x_1^2 + x_1^3)$$

by multiplying out

$$\begin{aligned}(x_2 - x_1)&(x_2^3 + x_2^2 x_1 + x_2 x_1^2 + x_1^3) \\ &= x_2^4 + x_2^3 x_1 + x_2^2 x_1^2 + x_2 x_1^3 - x_2^3 x_1 - x_2^2 x_1^2 - x_2 x_1^3 - x_1^4\end{aligned}$$

and then cancelling the terms in the middle to get $x_2^4 - x_1^4$.

This means that

$$|f(x_2) - f(x_1)| = |x_2^3 + x_2^2 x_1 + x_2 x_1^2 + x_1^3|\,|x_2 - x_1|.$$

We have the desired difference $|x_2 - x_1|$ on the right and we just have to bound the factor $|x_2^3 + x_2^2 x_1 + x_2 x_1^2 + x_1^3|$. By the triangle inequality

$$|x_2^3 + x_2^2 x_1 + x_2 x_1^2 + x_1^3| \le |x_2|^3 + |x_2|^2|x_1| + |x_2||x_1|^2 + |x_1|^3.$$

Now because x_1 and x_2 are in I, $|x_1| \le 2$ and $|x_2| \le 2$, so

$$|x_2^3 + x_2^2 x_1 + x_2 x_1^2 + x_1^3| \le 2^3 + 2^2\,2 + 2\,2^2 + 2^3 = 32$$

and

$$|f(x_2) - f(x_1)| \le 32|x_2 - x_1|.$$

Recall that the Lipschitz constant of $f(x) = x^2$ on I is $L_f = 4$. The fact that the Lipschitz constant of x^4 is larger than the constant for x^2 on $[-2, 2]$ is not surprising considering the plots of the two functions, see Fig. 10.12.

We can use the same technique to show that the function $f(x) = x^m$ is Lipschitz continuous where m is any natural number.

EXAMPLE 12.11. The function $f(x) = x^m$ is Lipschitz continuous on any interval $I = [-a, a]$, where a is a positive rational number, with Lipschitz constant $L_f = ma^{m-1}$. Given x_1 and x_2 in I, we want to estimate

$$|f(x_2) - f(x_1)| = |x_2^m - x_1^m|$$

in terms of $|x_2 - x_1|$. We can do this using the fact that

$$\begin{aligned} x_2^m - x_1^m &= (x_2 - x_1)(x_2^{m-1} + x_2^{m-2}x_1 + \cdots + x_2 x_1^{m-2} + x_1^{m-1}) \\ &= (x_2 - x_1) \sum_{i=0}^{m-1} x_2^{m-1-i} x_1^i. \end{aligned}$$

We show this by first multiplying out

$$(x_2 - x_1) \sum_{i=0}^{m-1} x_2^{m-1-i} x_1^i = \sum_{i=0}^{m-1} x_2^{m-i} x_1^i - \sum_{i=0}^{m-1} x_2^{m-1-i} x_1^{i+1}$$

To see that there is a lot of cancellation among the terms in the middle in the two sums on the right, we separate the first term out of the first sum and the last term in the second sum

$$(x_2 - x_1) \sum_{i=0}^{m-1} x_2^{m-1-i} x_1^i = x_2^m + \sum_{i=1}^{m-1} x_2^{m-i} x_1^i - \sum_{i=0}^{m-2} x_2^{m-1-i} x_1^{i+1} - x_1^m$$

and then changing the index in the second sum to get

$$\begin{aligned} (x_2 - x_1) &\sum_{i=0}^{m-1} x_2^{m-1-i} x_1^i \\ &= x_2^m + \sum_{i=1}^{m-1} x_2^{m-i} x_1^i - \sum_{i=1}^{m-1} x_2^{m-i} x_1^i - x_1^m = x_2^m - x_1^m. \end{aligned}$$

This is tedious, but it is good practice to go through the details and make sure this argument is correct.

This means that

$$|f(x_2) - f(x_1)| = \left| \sum_{i=0}^{m-1} x_2^{m-1-i} x_1^i \right| |x_2 - x_1|.$$

We have the desired difference $|x_2 - x_1|$ on the right and we just have to bound the factor

$$\left| \sum_{i=0}^{m-1} x_2^{m-1-i} x_1^i \right|.$$

By the triangle inequality

$$\left|\sum_{i=0}^{m-1} x_2^{m-1-i} x_1^i\right| \leq \sum_{i=0}^{m-1} |x_2|^{m-1-i} |x_1|^i.$$

Now because x_1 and x_2 are in $[-a, a]$, $|x_1| \leq a$ and $|x_2| \leq a$. So

$$\left|\sum_{i=0}^{m-1} x_2^{m-1-i} x_1^i\right| \leq \sum_{i=0}^{m-1} a^{m-1-i} a^i = \sum_{i=0}^{m-1} a^{m-1} = m a^{m-1}.$$

and

$$|f(x_2) - f(x_1)| \leq m a^{m-1} |x_2 - x_1|.$$

12.5 Linear combinations of functions

Now that we have seen that the monomials are Lipschitz continuous on a given bounded interval, it is a short step to show that any polynomial is Lipschitz continuous on a given bounded interval. But rather than just do this for polynomials, we show that a linear combination of arbitrary Lipschitz continuous functions is Lipschitz continuous

Suppose that f_1 is Lipschitz continuous with constant L_1 and f_2 is Lipschitz continuous with constant L_2 on the interval I. Note that here (and below) we condense the notation and write e.g. L_1 instead of L_{f_1}. Then $f_1 + f_2$ is Lipschitz continuous with constant $L_1 + L_2$ on I, because if we choose two points x and y in I, then

$$\begin{aligned}
|(f_1 + f_2)(y) - (f_1 + f_2)(x)| &= |(f_1(y) - f_1(x)) + (f_2(y) - f_2(x))| \\
&\leq |f_1(y) - f_1(x)| + |f_2(y) - f_2(x)| \\
&\leq L_1|y - x| + L_2|y - x| \\
&= (L_1 + L_2)|y - x|
\end{aligned}$$

by the triangle inequality. The same argument shows that $f_2 - f_1$ is Lipschitz continuous with constant $L_1 + L_2$ as well (not $L_1 - L_2$ of course!). It is even easier to show that if $f(x)$ is Lipschitz continuous on an interval I with Lipschitz constant L then $cf(x)$ is Lipschitz continuous on I with Lipschitz constant $|c|L$.

From these two facts, it is a short step to extend the result to any linear combination of Lipschitz continuous functions. Suppose that $f_1, \cdots, f_n$ are Lipschitz continuous on I with Lipschitz constants $L_1, \cdots, L_n$ respectively. We use induction, so we begin by considering the linear combination of two functions. From the remarks above, it follows that $c_1 f_1 + c_2 f_2$ is Lipschitz continuous with constant $|c_1|L_1 + |c_2|L_2$. Next given $i \leq n$, we assume that

$c_1f_1 + \cdots + c_{i-1}f_{i-1}$ is Lipschitz continuous with constant $|c_1|L_1 + \cdots + |c_{i-1}|L_{i-1}$. To prove the result for i, we write

$$c_1f_1 + \cdots + c_if_i = \big(c_1f_1 + \cdots + c_{i-1}f_{i-1}\big) + c_nf_n.$$

But the assumption on $\big(c_1f_1 + \cdots + c_{i-1}f_{i-1}\big)$ means that we have written $c_1f_1 + \cdots + c_if_i$ as the sum of two Lipschitz continuous functions, namely $\big(c_1f_1 + \cdots + c_{i-1}f_{i-1}\big)$ and c_nf_n. The result follows by the result for the linear combination of two functions. By induction, we have proved

Theorem 12.1 *Suppose that $f_1, \cdots, f_n$ are Lipschitz continuous on I with Lipschitz constants $L_1, \cdots, L_n$ respectively. Then the linear combination $c_1f_1+\cdots+c_nf_n$ is Lipschitz continuous on I with Lipschitz constant $|c_1|L_1+\cdots+|c_n|L_n$.*

Corollary 12.2 *A polynomial is Lipschitz continuous on any bounded interval.*

EXAMPLE 12.12. We show that the function $f(x) = x^4 - 3x^2$ is Lipschitz continuous on $[-2, 2]$, with constant $L_f = 44$. For x_1 and x_2 in $[-2, 2]$, we have to estimate

$$\begin{aligned}|f(x_2) - f(x_1)| &= |(x_2^4 - 3x_2^2) - (x_1^4 - 3x_1^2)| \\ &= |(x_2^4 - x_1^4) - (3x_2^2 - 3x_1^2)| \\ &\le |x_2^4 - x_1^4| + 3|x_2^2 - x_1^2|.\end{aligned}$$

From Example 12.11, we know that x^4 is Lipschitz continuous on $[-2, 2]$ with constant 32 while x^2 is Lipschitz continuous on $[-2, 2]$ with Lipschitz constant 4. Therefore

$$|f(x_2) - f(x_1)| \le 32|x_2 - x_1| + 3 \times 4|x_2 - x_1| = 44|x_2 - x_1|.$$

12.6 Bounded functions

Lipschitz continuity is related to another important property of a function called boundedness. A function f is *bounded* on a set of rational numbers I if there is a constant M such that, see Fig. 12.4

$$|f(x)| \le M \text{ for all } x \text{ in } I.$$

In fact if we think about the estimates we have made to verify the definition of Lipschitz continuity (12.2), we see that in every case these involved showing that some function is bounded on the given interval.

EXAMPLE 12.13. To show that $f(x) = x^2$ is Lipschitz continuous on $[-2, 2]$ in Example 12.4, we proved that $|x_1 + x_2| \le 4$ for x_1 and x_2 in $[-2, 2]$.

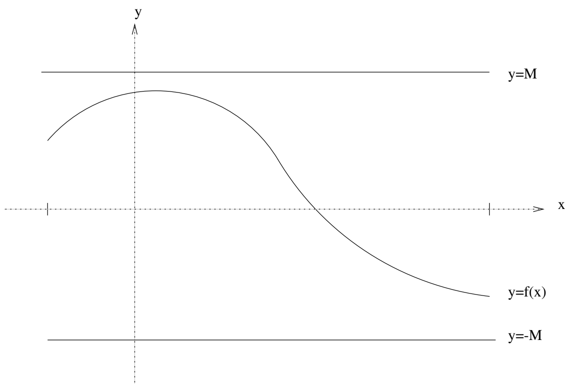


FIGURE 12.4. A bounded function on I.

It turns out that a function that is Lipschitz continuous on a bounded domain is automatically bounded on that domain. To be more precise, suppose that a function f is Lipschitz continuous with Lipschitz constant L_f on a bounded set I with size a and choose a point y in I. Then for any other point x in I

$$|f(x) - f(y)| \leq L_f|x - y|.$$

First we know that $|x - y| \leq |x| + |y| \leq 2a$. Also, since $|b + c| \leq |d|$ means that $|b| \leq |d| + |c|$ for any numbers a, b, c, we get

$$|f(x)| \leq |f(y)| + L_f|x - y| \leq |f(y)| + 2L_f a.$$

Even though we don't know $|f(y)|$, we do know that it is finite. This shows that $|f(x)|$ is bounded by the constant $M = |f(y)| + 2L_f a$ for any x in Q. We express this by saying that $f(x)$ is *bounded on* I. We have thus proved

Theorem 12.3 *A Lipschitz continuous function on a bounded set I is bounded on I.*

EXAMPLE 12.14. In Example 12.12, we showed that $f(x) = x^4 + 3x^2$ is Lipschitz continuous on $[-2, 2]$ with Lipschitz constant $L_f = 44$. Using this argument, we find that

$$|f(x)| \leq |f(0)| + 44|x - 0| \leq 0 + 44 \times 2 = 88$$

for any x in $[-2, 2]$. Since x^4 is increasing for $0 \leq x$, in fact we know that $|f(x)| \leq |f(2)| = 16$ for any x in $[-2, 2]$. So the estimate on the size of $|f|$ using the Lipschitz constant is not very accurate.

12.7 The product of functions

The next step in investigating which functions are Lipschitz continuous is to consider the product of two Lipschitz continuous functions on a bounded interval I. We show that the product is also Lipschitz continuous on I. More precisely, if f_1 is Lipschitz continuous with constant L_1 and f_2 is Lipschitz continuous with constant L_2 on a bounded interval I then $f_1 f_2$ is Lipschitz continuous on I. We choose two points x and y in I and estimate by using the old trick of adding and subtracting the same quantity

$$\begin{aligned}
&|f_1(y)f_2(y) - f_1(x)f_2(x)| \\
&\qquad = |f_1(y)f_2(y) - f_1(y)f_2(x) + f_1(y)f_2(x) - f_1(x)f_2(x)| \\
&\qquad \leq |f_1(y)f_2(y) - f_1(y)f_2(x)| + |f_1(y)f_2(x) - f_1(x)f_2(x)| \\
&\qquad = |f_1(y)|\,|f_2(y) - ff_2(x)| + |f_2(x)|\,|f_1(y) - f_1(x)|
\end{aligned}$$

Now Theorem 12.3, which says that Lipschitz continuous functions are bounded, implies there is some constant M such that $|f_1(y)| \leq M$ and $|f_2(x)| \leq M$ for $x, y \in I$. Using the Lipschitz continuity of f_1 and f_2 in I, we find

$$\begin{aligned}
|f_1(y)f_2(y) - f_1(x)f_2(x)| &\leq ML_1|y-x| + ML_2|y-x| \\
&= M(L_1 + L_2)|y-x|.
\end{aligned}$$

We summarize

Theorem 12.4 *If f_1 and f_2 are Lipschitz continuous on a bounded interval I then $f_1 f_2$ is Lipschitz continuous on I.*

EXAMPLE 12.15. The function $f(x) = (x^2+5)^{10}$ is Lipschitz continuous on the set $I = [-10, 10]$ because $x^2 + 5$ is Lipschitz continuous on I and therefore $(x^2+5)^{10} = (x^2+5)(x^2+5)\cdots(x^2+5)$ is as well by Theorem 12.4.

12.8 The quotient of functions

Continuing our investigation, we now consider the ratio of two Lipschitz continuous functions. In this case however, we require more information about the function in the denominator than just that it is Lipschitz continuous. We also have to know that it does not become too small. To understand this, we first consider an example.

EXAMPLE 12.16. We show that $f(x) = 1/x^2$ is Lipschitz continuous on the interval $[1/2, 2]$, with Lipschitz constant $L = 64$. We choose two points x_1 and x_2 in Q and we estimate the change

$$|f(x_2) - f(x_1)| = \left|\frac{1}{x_2^2} - \frac{1}{x_1^2}\right|$$

by first doing some algebra

$$\frac{1}{x_2^2} - \frac{1}{x_1^2} = \frac{x_1^2}{x_1^2 x_2^2} - \frac{x_2^2}{x_1^2 x_2^2} = \frac{x_1^2 - x_2^2}{x_1^2 x_2^2} = \frac{(x_1 + x_2)(x_1 - x_2)}{x_1^2 x_2^2}.$$

This means that

$$|f(x_2) - f(x_1)| = \left| \frac{x_1 + x_2}{x_1^2 x_2^2} \right| |x_2 - x_1|.$$

Now we have the good difference on the right, we just have to bound the factor. The numerator of the factor is the same as in Example 12.4, and we know that

$$|x_1 + x_2| \leq 4.$$

We also know that

$$x_1 \geq \frac{1}{2} \text{ implies } \frac{1}{x_1} \leq 2 \text{ implies } \frac{1}{x_1^2} \leq 4$$

and likewise $\frac{1}{x_2^2} \leq 4$. So we get

$$|f(x_2) - f(x_1)| \leq 4 \times 4 \times 4\,|x_2 - x_1| = 64|x_2 - x_1|.$$

In this example, we have to use the fact that the left-hand endpoint of the interval I is $1/2$. The closer the left-hand endpoint is to zero, the larger the Lipschitz constant will be. In fact, $1/x^2$ is **not** Lipschitz continuous on $[0, 2]$.

We mimic this example in the general case f_1/f_2 by assuming that the denominator f_2 is *bounded below* by a positive constant. We give the proof of the following theorem as an exercise.

Theorem 12.5 *Assume that f_1 and f_2 are Lipschitz continuous on a bounded set I with constants L_1 and L_2 and moreover assume there is a constant $m > 0$ such that $|f_2(x)| \geq m$ for all x in I. Then f_1/f_2 is Lipschitz continuous on I.*

EXAMPLE 12.17. The function $1/x^2$ does not satisfy the assumptions of Theorem 12.5 on the interval $[0, 2]$ and we know that it is not Lipschitz continuous on that interval.

12.9 The composition of functions

We conclude the investigation into Lipschitz continuity by considering the composition of Lipschitz continuous functions. This is actually easier than either products or ratios of functions. The only complication is that we have to be careful about the domains and ranges of the functions. Consider the composition $f_2(f_1(x))$. Presumably, we have to restrict x to an interval on which f_1 is Lipschitz continuous and we also have to make sure that the values of f_1 are in a set on which f_2 is Lipschitz continuous.

So we assume that f_1 is Lipschitz continuous on I_1 with constant L_1 and that f_2 is Lipschitz continuous on I_2 with constant L_2. If x and y are points in I_1 then as long as $f_1(x)$ and $f_1(y)$ are in I_2 then

$$|f_2(f_1(y)) - f_2(f_1(x))| \leq L_2|f_1(y) - f_1(x)| \leq L_1 L_2 |y - x|.$$

We summarize as a theorem.

Theorem 12.6 *Let f_1 be Lipschitz continuous on a set I_1 with Lipschitz constant L_1 and f_2 be Lipschitz continuous on I_2 with Lipschitz constant L_2 such that $f_1(I_1) \subset I_2$. Then the composite function $= f_2 \circ f_1$ is Lipschitz continuous on I_1 with Lipschitz constant $L_1 L_2$.*

EXAMPLE 12.18. The function $f(x) = (2x-1)^4$ is Lipschitz continuous on any bounded interval since $f_1(x) = 2x - 1$ and $f_2(x) = x^4$ are Lipschitz continuous on any bounded interval. If we consider the interval $[-.5, 1.5]$ then $f_1(I) \subset [-2, 2]$. From Example 12.10, we know that x^4 is Lipschitz continuous on $[-2, 2]$ with Lipschitz constant 32 while the Lipschitz constant of $2x - 1$ is 2. Therefore, f is Lipschitz continuous on $[-.5, 1.5]$ with constant 64.

EXAMPLE 12.19. The function $1/(x^2 - 4)$ is Lipschitz continuous on any closed interval that does not contain either 2 or -2. This follows because $f_1(x) = x^2 - 4$ is Lipschitz continuous on any bounded interval while $f_2(x) = 1/x$ is Lipschitz continuous on any closed interval that does not contain 0. To avoid zero, we must avoid $x^2 = 4$ or $x = \pm 2$.

12.10 Functions of two rational variables

Until now, we have considered functions $f(x)$ of one rational variable x. But of course, there are functions that depend on more than one input. Consider for example the function

$$f(x_1, x_2) = x_1 + x_2,$$

which to each pair of rational numbers x_1 and x_2 associates the sum $x_1 + x_2$. We may write this as $f : \mathbb{Q} \times \mathbb{Q} \to \mathbb{Q}$, meaning that to each $x_1 \in Q$ and $x_2 \in \mathbb{Q}$ we associate a value $f(x_1, x_2) \in \mathbb{Q}$. For example, $f(x_1, x_2) = x_1 + x_2$. We say that $f(x_1, x_2)$ is a *function of two independent rational variables x_1 and x_2*. Here, we think of $\mathbb{Q} \times \mathbb{Q}$ as the set of all pairs (x_1, x_2) with $x_1 \in \mathbb{Q}$ and $x_2 \in \mathbb{Q}$.

We shall write $\mathbb{Q}^2 = \mathbb{Q}\times\mathbb{Q}$ and consider $f(x_1, x_2) = x_1 + x_2$ as a function $f : \mathbb{Q}^2 \to \mathbb{Q}$. We will also consider functions $f : I \times J \to Q$, where I and J are subsets such as intervals, of Q. This just means that for each $x_1 \in I$ and $x_2 \in J$, we associate a value $f(x_1, x_2) \in \mathbb{Q}$.

We may naturally extend the concept of Lipschitz continuity to functions of two rational variables. We say that $f : I\times J \to \mathbb{Q}$ is Lipschitz continuous with Lipschitz constant L_f if

$$|f(x_1, y_1) - f(x_2, y_2)| \le L_f(|x_1 - x_2| + |y_1 - y_2|)$$

for $x_1, x_2 \in I$ and $y_1, y_2 \in J$.

Example 12.20. The function $f : \mathbb{Q}^2 \to \mathbb{Q}$ defined by $f(x_1, x_2) = x_1 + x_2$ is Lipschitz continuous with Lipschitz constant $L_f = 1$.

Example 12.21. The function $f : [0, 2] \times [0, 2] \to \mathbb{Q}$ defined by $f(x_1, x_2) = x_1x_2$ is Lipschitz continuous with Lipschitz constant $L_f = 2$, since for $x_1, x_2 \in [0, 1]$

$$\begin{aligned} |x_1x_2 - y_1y_2| &= |x_1x_2 - y_1x_2 + y_1x_2 - y_1y_2| \\ &\le |x_1 - y_1|x_2 + y_1|x_2 - y_2| \le 2(|x_1 - y_1| + |x_2 - y_2|). \end{aligned}$$

12.11 Functions of several rational variables

The concept of a function also extends to several variables, i.e. we consider functions $f(x_1, ..., x_d)$ of d rational variables. We write $f : \mathbb{R}^d \to \mathbb{Q}$ if for given rational numbers $x_1, \cdots, x_d$, a rational number denoted by $f(x_1, ..., x_d)$ is given.

The definition of Lipschitz continuity also directly extends. We say that $f : \mathbb{Q}^d \to \mathbb{Q}$ is Lipschitz continuous with Lipschitz constant L_f if for all $x_1, \cdots, x_d \in \mathbb{Q}$ and $y_1, \cdots, y_d \in \mathbb{Q}$,

$$|f(x_1, ..., x_d) - f(y_1, ..., y_d)| \le L_f(|x_1 - y_1| + \cdots + |x_d - y_d|).$$

Example 12.22. The function $f : \mathbb{R}^d \to \mathbb{Q}$ defined by $f(x_1, ..., x_d) = x_1 + x_2 + \cdots x_d$ is Lipschitz continuous with Lipschitz constant $L_f = 1$.

Chapter 12 Problems

12.1. Verify the claims in Example 12.7.

12.2. Show that $f(x) = x^2$ is Lipschitz continuous on $[10, 13]$ directly and compute a Lipschitz constant.

12.3. Show that $f(x) = 4x - 2x^2$ is Lipschitz continuous on $[-2, 2]$ directly and compute a Lipschitz constant.

12.4. Show that $f(x) = x^3$ is Lipschitz continuous on $[-2, 2]$ directly and compute a Lipschitz constant.

12.5. Show that $f(x) = |x|$ is Lipschitz continuous on $\mathbb{Q}$ directly and compute a Lipschitz constant.

12.6. In Example 12.10, we show that x^4 is Lipschitz continuous on $[-2, 2]$ with Lipschitz constant $L = 32$. Explain why this is a reasonable value for the Lipschitz constant.

12.7. Show that $f(x) = 1/x^2$ is Lipschitz continuous on $[1, 2]$ directly and compute the Lipschitz constant.

12.8. Show that $f(x) = 1/(x^2 + 1)$ is Lipschitz continuous on $[-2, 2]$ directly and compute a Lipschitz constant.

12.9. Compute the Lipschitz constant of $f(x) = 1/x$ on the intervals (a) $[.1, 1]$, (b) $[.01, 1]$, and $[.001, 1]$.

12.10. Find the Lipschitz constant of the function $f(x) = \sqrt{x}$ with $D(f) = (\delta, \infty)$ for given $\delta > 0$.

12.11. Explain why $f(x) = 1/x$ is not Lipschitz continuous on $(0, 1]$.

12.12. (a) Explain why the function

$$f(x) = \begin{cases} 1, & x < 0 \\ x^2, & x \geq 0 \end{cases}$$

is **not** Lipschitz continuous on $[-1, 1]$. (b) Is f Lipschitz continuous on $[1, 4]$?

12.13. Suppose the Lipschitz constant L of a function f is equal to $L = 10^{100}$. Discuss the continuity properties of $f(x)$ and in particular decide if f continuous from a practical point of view.

12.14. Assume that f_1 is Lipschitz continuous with constant L_1, f_2 is Lipschitz continuous with constant L_2 on a set I, and c is a number. Show that $f_1 - f_2$ is Lipschitz continuous with constant $L_1 + L_2$ on I and cf_1 is Lipschitz continuous with constant cL_1 on I.

12.15. Show that the Lipschitz constant of a polynomial $f(x) = \sum_{i=0}^{n} a_i x^i$ on the interval $[-c, c]$ is

$$L = \sum_{i=1}^{n} |a_i| i c^{i-1} = |a_1| + 2c|a_2| + \cdots + nc^{n-1}|a_n|.$$

12.16. Explain why $f(x) = 1/x$ is not bounded on $[-1, 0]$.

12.17. Prove Theorem 12.5.

12.18. Use the theorems in this chapter to show that the following functions are Lipschitz continuous on the given intervals and try to estimate a Lipschitz constant or prove they are not Lipschitz continuous.

(a) $f(x) = 2x^4 - 16x^2 + 5x$ on $[-2, 2]$

(b) $\dfrac{1}{x^2 - 1}$ on $\left[-\dfrac{1}{2}, \dfrac{1}{2}\right]$

(c) $\dfrac{1}{x^2 - 2x - 3}$ on $[2, 3)$

(d) $\left(1 + \dfrac{1}{x}\right)^4$ on $[1, 2]$

12.19. Show the function

$$f(x) = \frac{1}{c_1 x + c_2(1 - x)}$$

where $c_1 > 0$ and $c_2 > 0$ is Lipschitz continuous on $[0, 1]$.

13
Sequences and limits

He sat down and thought, in the most thoughtful way he could think.
(Winnie-the-Pooh)

13.1 A first encounter with sequences and limits

The decimal expansions of rational numbers discussed in chapter Rational Numbers leads into the concepts *sequence*, *converging sequence* and *limit* of a sequence, which play a fundamental role in mathematics. The development of calculus has largely been a struggle to come to grips with certain evasive aspects of these concepts. We will try to uncover the mysteries by being as concrete and down-to-earth as possible.

We begin recalling the decimal expansion 1.11... of $\frac{10}{9}$, and that by (7.7)

$$\frac{10}{9} = 1.11\cdots 11_n + \frac{1}{9}10^{-n}. \tag{13.1}$$

Rewriting this equation and replacing for simplicity $\frac{1}{9}10^{-n}$ by the upper bound 10^{-n}, we get the following *estimate* for the difference between $1.111\cdots 11_n$ and $10/9$,

$$\left|\frac{10}{9} - 1.11\cdots 11_n\right| \leq 10^{-n}. \tag{13.2}$$

This estimate shows that we may consider $1.11\cdots 11_n$ as an approximation of $10/9$, which becomes increasingly accurate as the number of decimal places n increases. In other words, the *error* $|10/9 - 1.11\cdots 11_n|$ can be made as small as we please by taking n sufficiently large. If we want the error to be smaller than or equal to 10^{-10}, then we simply choose $n \geq 10$.

We may view the successive approximations 1.1, 1.11, 1.111, $1.11...11_n$, and so on, as a *sequence* of numbers a_n, with $n = 1, 2, 3, ...$, where $a_1 = 1.1$, $a_2 = 1.11$,..., $a_n = 1.11...11_n$,..., are called the *elements of the sequence*. More generally, a sequence a_1, a_2, a_3,..., is a never-ending list of elements a_1, a_2, a_3,..., where the index takes successively the values of the natural numbers $1, 2, 3, ...$ A *sequence of rational numbers* is a list a_1, a_2, a_3,..., where each element a_n is a rational number. We will denote a sequence by

$$\{a_n\}_{n=1}^{\infty}$$

which thus means the neverending list a_1, a_2, a_3,..., of elements a_n, with the index n going through the natural numbers $n = 1, 2, 3, ...$ The symbol ∞, called "infinity", indicates that the list continues for ever in the same sense that the natural numbers $1, 2, 3, ...,$ continues for ever without coming to an end.

We now return to the sequence of rational numbers $\{a_n\}_{n=1}^{\infty}$, where $a_n = 1.11..11_n$, that is the sequence $\{1.11..11_n\}_{n=1}^{\infty}$. The accuracy of element $a_n = 1.11...11_n$, as an approximation of $\frac{10}{9}$, increases as the number of decimals n increases. Each number in the sequence in turn is a better approximation to 10/9 than the preceding number and as we move from left to right the numbers become ever closer to 10/9. An advantage of considering the sequence $\{1.11..11_n\}_{n=1}^{\infty}$ or neverending list 1.1, 1.11, 1.111,..., is that we are ready to meet any accuracy requirement that could be posed. If we just consider one element, say $1.11..11_{10}$, then we could not meet an accuracy requirement in the approximation of $\frac{10}{9}$ of say 10^{-15}. But if we have the whole sequence at hand, then we can pick the element $1.11..11_{16}$ or $1.11..11_{17}$ or more generally any $1.11..11_n$ with $n \geq 15$, as a decimal approximation of $\frac{10}{9}$ with an error less than 10^{-15}. The sequence thus gives us a whole "bag" of numbers, or a collection of approximations with which we can meet any accuracy requirement in the approximation of $\frac{10}{9}$. The sequence $1.1, ..., 1.11..11_n, ...,$ thus can be viewed as a collection of successively more accurate approximations of $\frac{10}{9}$, where we can satisfy any desired accuracy.

We say that the sequence $\{1.11..11_n\}_{n=1}^{\infty}$ *converges* to the value $\frac{10}{9}$, since the difference between $\frac{10}{9}$ and $1.11..11_n$ becomes smaller than any given positive number if only we take n large enough, as follows from (13.2). We say that $\frac{10}{9}$ is the *limit* of the sequence $\{a_n\}_{n=1}^{\infty}$=$\{1, 11..11_n\}_{n=1}^{\infty}$. We will express the convergence of the sequence $\{a_n\}_{n=1}^{\infty}$ with elements $a_n = 1.11...11_n$, as follows:

$$\lim_{n\to\infty} a_n = \frac{10}{9} \quad \text{or} \quad \lim_{n\to\infty} 1.11..11_n = \frac{10}{9}.$$

The limit $\frac{10}{9}$ does not have a finite decimal expansion. The elements $1.11..11_n$ of the converging sequence $\{1.11..11_n\}_{n=1}^{\infty}$ are finite decimal approximations of the limit $\frac{10}{9}$, with an error which is smaller than any given positive number if we only take n large enough.

Suppose that we restrict ourselves to work with finite decimal expansions, which is what a computer usually does. In this case we cannot exactly express the value $\frac{10}{9}$ with the available resources, because $\frac{10}{9}$ does not have a finite decimal expansion. As a substitute or approximation we may choose for example $1.11..11_{10}$, but there is limit to the accuracy with this single element. It would not be entirely correct to say that $\frac{10}{9} = 1.11..11_{10}$. If we instead have the whole sequence $\{1.11..11_n\}_{n=1}^{\infty}$ at hand, then we can meet any accuracy by choosing the element $1.11..11_n$ with n large enough. Choosing more and more decimals, we could increase the accuracy to any desired degree.

The sequence $\{1.11..11_n\}_{n=1}^{\infty}$ includes finite decimal approximations of $\frac{10}{9}$ satisfying any given positive tolerance or accuracy requirement. This is sometimes expressed as

$$1.111... = \frac{10}{9},$$

where the three little dots are there to indicate that any precision could be attained by taking sufficiently many decimals (all equal to 1). Another way of writing this, would be

$$\lim_{n\to\infty} 1.11..11_n = \frac{10}{9},$$

avoiding the possible ambiguity using the three little dots.

13.2 Socket wrench sets

To tighten or loosen a hex bolt with head diameter $2/3$, a mechanic needs to use a socket wrench of a slightly bigger size. The tolerance on the difference between the sizes of the bolt and the wrench depend on the tightness, the material of the bolt and the wrench, and conditions such as whether the bolt threads are lubricated and whether the bolt is rusty or not. If the wrench is too large then the head of the bolt will simply be stripped before the bolt is tightened or loosened. We show two wrenches with different tolerances in Fig. 13.1.

An *amateur mechanic* would have one socket, of say dimension 0.7. A *pro mechanic* would perhaps have 10 sockets of dimensions 0.7, 0.67, 0.667,....,$0.66\cdots667_{10}$. Both the amateur and pro would get stuck under sufficiently tough conditions because the socket would be too large to do the job.

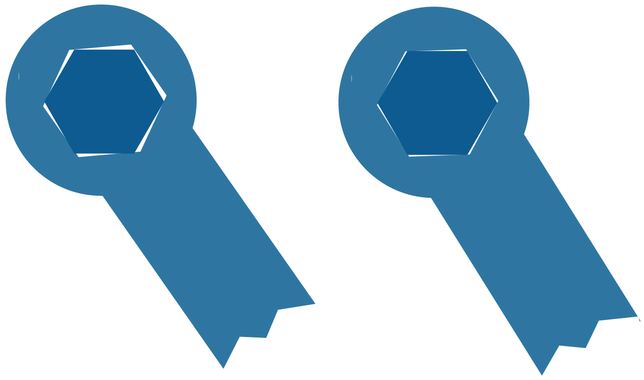


FIGURE 13.1. Two socket wrenches with different tolerances.

A *ideal expert mechanic* would have the whole sequence $\{0.66\cdots667_n\}_{n=1}^{\infty}$ at his/her disposal with the error of wrench number n being estimated by

$$|0.66\cdots667_n - \frac{2}{3}| \leq 10^{-n}.$$

The ideal expert can thus reach into the tool chest and pull out a wrench that meets any accuracy requirement, and would thus be able to turn the bolt under arbitrarily tough conditions, or meet any crank torque specified by a bicycle manufacturer. More precisely, the ideal expert could be thought of as being able to *construct* a socket himself to meet any given tolerance or accuracy. If necessary, the ideal mechanic could *construct* a wrench of for example the dimension $0.66\cdots67_{20}$, unless he already has such a wrench in his (big) tool chest. The amateur and pro mechanic would not have this capability of constructing their own wrenches, but would have to be content with their ready-made wrench sets (which they could buy in the hard-ware store). We expect the cost to construct a wrench of dimension $0.6\cdots67_n$ to increase (rapidly) with n, since the precision in the construction process has to improve.

As a general point, computing the numbers $0.66\cdots67_n$ by long division of $\frac{2}{3}$, requires more work as n increases. What we gain from doing more work is better accuracy in using $0.66\cdots67_n$ as an approximation to $\frac{2}{3}$. Trading work for accuracy is the central idea behind solving equations using computation, especially on a computer. An estimate like (13.2) gives a quantitative measurement of how much accuracy we gain for each increase in work and so such estimates are useful not only to mathematicians but to engineers and scientists.

The need of approximating better and better in this case may be seen as an incompatibility of two systems: the bolt has dimension $\frac{2}{3}$ in the system of rational numbers, and the wrenches come in the decimal system 0.7, 0.67, 0.667,... and there is no wrench of size exactly $\frac{2}{3}$.

13.3 J.P. Johansson's adjustable wrenches

The adjustable wrench is a Swedish invention created 1891 by the genius J.P. Johansson (1839-1924) see Fig. 13.2. In principle the adjustable wrench is an analog device which fits a bolt of any size within a certain range. Every mechanic knows that an adjustable wrench may fail in cases when a properly chosen fixed size wrench does not, because the size of the adjustable wrench is not completely stable under increasing torque. .

FIGURE 13.2. The Swedish inventor J.P.Johansson with two adjustable wrenches of different design.

13.4 The power of language: from infinitely many to one

The decimal expansion 0.6666.... of $\frac{2}{3}$ contains infinitely many decimals. The sequence $\{0.66\cdots667_n\}_{n=1}^{\infty}$ contains infinitely many elements, which are increasingly accurate approximations of $\frac{2}{3}$. Talking or thinking of infinitely many decimals or infinitely many elements, presents a serious difficulty, which is handled by introducing the concept of a sequence. A sequence has *infinitely many elements*, but the sequence itself is just *one entity.* We thus group the infinitely many elements together to form one sequence, and thus pass from infinity to one. After this semantic construction, we are thus able to speak about *one* sequence and may momentarily forget that the sequence in fact has infinitely many elements.

This would be like speaking about the expert mechanics tool chest containing the sequence $\{0.66\cdots 667_n\}_{n=1}^{\infty}$ of infinitely many wrenches as one entity. One tool chest with infinitely many wrenches. To call a tool chest a wrench seems strange initially, but we could get there by first calling the tool chest something like a "super-wrench", and then later omit the "super".

Analogously, we could say that 0.6666.... is a "super-number" because it has infinitely many decimals, and then forget the "super" and say that 0.6666...., is a number. In fact, this makes complete sense since we identify 0.6666.... with $\frac{2}{3}$, which is a number. Below, we shall meet non-periodic infinite decimal expansions that do not correspond to rational numbers. Initially, we may think of these as some kind of "super-number", and then later will refer such numbers as "real numbers".

The discussion illustrates the usefulness of the concept of *one* set or sequence with *infinitely* many elements. Of course, we should be aware of the risk involved using the language to hide real facts. Political language is often used this way, which is one reason for the eroding credibility of politicians. As mathematicians, there is no reason that we should try to be as honest as possible, and use the language as clearly as possible.

13.5 The $\epsilon - N$ definition of a limit

The mathematical formulation of the idea of a limit says that the terms a_n of a convergent sequence $\{a_n\}_{n=1}^{\infty}$ differ from the limit A with as little as we please if only the index n is large enough, and we decided to write this as

$$\lim_{n\to\infty} a_n = A,$$

There is a mathematical jargon to express this fact that has become extremely popular. It was developed by Karl Weierstrass (1815-97), see Fig. 82.25 and takes the following form: The limit of the sequence $\{a_n\}_{n=1}^{\infty}$ equals A, which we write as

$$\lim_{n\to\infty} a_n = A,$$

if for any (rational) $\epsilon > 0$ there is a natural number N such that

$$|a_n - A| \leq \epsilon \quad \text{for all} \quad n \geq N$$

For example, we know that the value of $10/9$ is approximated by the element $1.11\cdots 1_n$ from the sequence $\{1.11\cdots 1_n\}$ to any specified accuracy (bigger than zero) by taking n sufficiently large. We know from (13.2) that

$$|\frac{10}{9} - 1.11\cdots 1_n| \leq 10^{-n},$$

and thus

$$|\frac{10}{9} - 1.11\cdots 1_n| \leq \epsilon$$

if $10^{-n} \leq \epsilon$. We can phrase this as

$$|\frac{10}{9} - 1.11\cdots 1_n| \leq \epsilon$$

if $n \geq N$, where $10^{-N} \leq \epsilon$. If $\epsilon = .p_1p_2\cdots$, where $p_1 = p_2 = \cdots = p_m = 0$, while $p_{m+1} \neq 0$, then we may choose any N such that $N \geq m$. We see that choosing ϵ smaller, requires N to be bigger, and thus N depends on ϵ.

We emphasize that the $\epsilon - N$ definition of convergence is a fancy way of saying that the difference $|A - a_n|$ can be made smaller than any given positive number if only n is taken large enough.

There is a risk (and temptation) in using the $\epsilon - N$ definition of convergence, instead of the more pedestrian "as small as we please if only n is large enough". The statement "$|A-a_n|$ can be made smaller than any given positive number if only n is large enough" is a very qualitative statement. Nothing is said about *how large* n has to be to reach a certain accuracy. A very qualitative statement is necessarily a bit vague. On the other hand, the statement "for any $\epsilon > 0$ there is an N such that $|A - a_n| \leq \epsilon$ if $n \geq N$" has the form of a very exact and precise statement, while in fact it may be as qualitative as the first statement, unless the dependence of N on ϵ is made clear. The risk is thus that using the $\epsilon - N$-jargon, we may get confused and believe that something vague, in fact is very precise. Of course there is also a temptation in this, which relates to the general idea of mathematics as something being extremely precise. So be cautious and don't get fooled by simple tricks: the $\epsilon - N$ limit definition is vague to the extent the dependence of N on ϵ is vague.

The concept of a limit of a sequence of numbers is central to calculus. It is closely connected to never-ending decimal expansions, that is decimal expansions with infinitely many non-zero decimals. The elements in the sequence with this connection are obtained by successively taking more and more decimals into account. In fact, the fundamental reason for looking at sequences comes form this connection. However, as happens, the idea of a sequence and limit has taken on a life of its own, which has been plaguing many students of calculus. We will try to refrain from excesses in this direction and keep a strong connection with the original motivation for introducing the concepts of sequences and limits, namely describing successively better and better approximations of solutions of equations.

FIGURE 13.3. Weierstrass to Sonya Kovalevskaya: "...dreamed and been enraptured of so many riddles that remain for us to solve, on finite and infinite spaces, on the stability of the world system, and on all the other major problems of the mathematics and the physics of the future. ... you have been close ...throughout my entire life ... and never have I found anyone who could bring me such understanding of the highest aims of science and such joyful accord with my intentions and basic principles as you".

We shall now practice the $\epsilon - N$ jargon in a couple of examples to show that certain sequences have limits. The sequences we present are "artificial", that is given by cooked-up formulas, but we use them to illustrate basic aspects. After going through these examples, the reader should be able to look through the apparent mystery of the $\epsilon - N$ definition, and understand that it expresses something intuitively quite simple. But remember: the $\epsilon - N$ definition of a limit is vague to the extent that the dependence of N on ϵ is vague.

EXAMPLE 13.1. The limit of the sequence $\{\frac{1}{n}\}_{n=1}^{\infty}$ equals 0, i.e.

$$\lim_{n\to\infty} \frac{1}{n} = 0.$$

Note that this is obvious simply because $\frac{1}{n}$ can be made as close to 0 as we please by taking n large enough. We shall now phrase this obvious (and trivial fact) using the $\epsilon - N$ jargon. We thus have to satisfy the devious mathematician who gives an $\epsilon > 0$ and asks for a natural number N such that

$$\left|\frac{1}{n} - 0\right| \leq \epsilon \tag{13.3}$$

for all $n \geq N$. Well, to satisfy this request we choose N to be any natural number larger than (or equal to) $1/\epsilon$, for instance the smallest natural number larger than or equal to $1/\epsilon$. Then (13.3) holds for $n \geq N$, and we have satisfied the devious demand, which shows that $\lim_{n\to\infty} \frac{1}{n} = 0$. In this example, the connection between ϵ and N is very clear: we can take N to be the smallest natural number larger than or equal to $1/\epsilon$. For example, if $\epsilon = 1/100$, then $N = 100$. We hope the reader can make the connection of the simple idea that $1/n$ gets as close to 0 as we please by taking n sufficiently large, and the more pompous phrasing of this idea in the $\epsilon - N$-jargon.

EXAMPLE 13.2. We next show that the limit of the sequence $\{\frac{n}{n+1}\}_{n=1}^{\infty} = \{\frac{1}{2}, \frac{2}{3}, \cdots\}$ equals 1, that is

$$\lim_{n\to\infty} \frac{n}{n+1} = 1. \tag{13.4}$$

We compute

$$\left|1 - \frac{n}{n+1}\right| = \left|\frac{n+1-n}{n+1}\right| = \frac{1}{n+1},$$

which shows that $\frac{n}{n+1}$ is arbitrarily close to 1 if n is large enough, and thus proves the claim. We now phrase this using the $\epsilon - N$ jargon. Let thus $\epsilon > 0$ be given. Now $\frac{1}{n+1} \leq \epsilon$ provided that $n \geq 1/\epsilon - 1$. Hence $\left|1 - \frac{n}{n+1}\right| \leq \epsilon$ for all $n \geq N$ provided N is chosen so that $N \geq 1/\epsilon - 1$. Again this proves the claim.

EXAMPLE 13.3. The sum

$$1 + r + r^2 + \cdots + r^n = \sum_{i=0}^{n} r^i = s_n$$

is said to be a *finite geometric series of order n with factor r*, including the powers r^i of the factor r up to $i = n$. We considered this series above with $r = 0.1$ and $s_n = 1.11\cdots 11_n$. We now consider an arbitrary value of the factor r in the range $|r| < 1$. We recall the formula

$$s_n = \sum_{i=0}^{n} r^i = \frac{1 - r^{n+1}}{1 - r}$$

valid for any $r \neq 1$. What happens as the number n of terms get bigger and bigger? To answer this it is natural to consider the sequence $\{s_n\}_{n=1}^{\infty}$. We shall prove that if $|r| < 1$, then

$$\lim_{n\to\infty} s_n = \lim_{n\to\infty} (1 + r + r^2 + \cdots + r^n) = \frac{1}{1 - r}, \tag{13.5}$$

which we will write as

$$\sum_{i=0}^{\infty} r^i = \frac{1}{1 - r} \quad \text{if } |r| < 1.$$

Intuitively, we feel that this is correct, because r^{n+1} gets as small as we please by taking n large enough (remember that $|r| < 1$). We say that $\sum_{i=0}^{\infty} r^i$ is an *infinite geometric series with factor* r.

We now give an $\epsilon - N$ proof of (13.5). We need to show that for any $\epsilon > 0$, there is an N such that

$$\left| \frac{1 - r^{n+1}}{1 - r} - \frac{1}{1 - r} \right| = \left| \frac{r^{n+1}}{1 - r} \right| \leq \epsilon$$

for all $n \geq N$. To this end it is sufficient, since $|r| < 1$, to find N such that

$$|r|^{N+1} \leq \epsilon \, |1 - r| \tag{13.6}$$

Since $|r| < 1$, we can make $|r|^{N+1}$ as small as we please by taking N sufficiently large, and thus we can also satisfy the inequality (13.6) by taking N sufficiently large. Below, we will define a function called the logarithm that we can use to get a precise value for N as a function of ϵ from (13.6).

13.6 A converging sequence has a unique limit

The limit of a converging sequence is uniquely defined. This should be self-evident from the fact that it is impossible to be arbitrarily close to two different numbers at the same time. Try! We now also give a more lengthy proof using a type of argument often found in math books. The reader could profit from going through this argument and understanding that something seemingly difficult, in fact can hide a very simple idea.

We start from the following variation of the *triangle inequality*, see Problem 7.15,

$$|a - b| \leq |a - c| + |c - b| \tag{13.7}$$

which holds for all a, b, and c. Suppose that the sequence $\{a_n\}_{n=1}^{\infty}$ converges to two possibly different numbers A_1 and A_2. Using (13.7) with $a = A_1$, $b = A_2$, and $c = a_n$, we get

$$|A_1 - A_2| \leq |a_n - A_1| + |a_n - A_2|$$

for any n. Now because a_n converges to A_1, we can make $|a_n - A_1|$ as small as we like, and in particular smaller than $\frac{1}{4}|A_1 - A_2|$ if $A_1 \neq A_2$, by taking n large enough. Likewise we can make $|a_n - A_2| \leq \frac{1}{4}|A_1 - A_2|$ by taking n large enough. By (13.7), this means that $|A_1 - A_2| \leq \frac{1}{2}|A_1 - A_2|$ for n large, which can only hold if $A_1 = A_2$, and thus contradicts the obstructional assumption $A_1 \neq A_2$, which therefore must be rejected.

We note that if $\lim_{n\to\infty} a_n = A$ then also, for example, $\lim_{n\to\infty} a_{n+1} = A$, and $\lim_{n\to\infty} a_{n+7} = A$. In other word, only the "very tail" of a sequence $\{a_n\}$ matters to the limit $\lim_{n\to\infty} a_n$.

13.7 Lipschitz continuous functions and converging sequences

A basic reason for introducing the concept of a Lipschitz continuous function $f : \mathbb{Q} \to \mathbb{Q}$ is its relation to sequences of rational numbers. The fundamental issue is the following. Let $\{a_n\}$ be a converging sequence with rational limit $\lim_{n\to\infty} a_n$ and let $f : \mathbb{Q} \to \mathbb{Q}$ be a Lipschitz continuous function with Lipschitz constant L. What can be said about the sequence $\{f(a_n)\}$? Does it converge and if so, to what?

The answer is easy to state: the sequence $\{f(a_n)\}$ converges and

$$\lim_{n\to\infty} f(a_n) = f(\lim_{n\to\infty} a_n).$$

The proof is also easy. By the Lipschitz continuity of $f : \mathbb{Q} \to \mathbb{Q}$, we have

$$|f(a_m) - f(\lim_{n\to\infty} a_n)| \leq L|a_m - \lim_{n\to\infty} a_n|.$$

Since $\{a_n\}$ converges to A, the right-hand side can be made smaller than any given positive number by taking m large enough, and thus we can also make the left hand side smaller than any positive number by choosing m large enough, which shows the desired result.

Note that since $\lim_{n\to\infty} a_n$ is a rational number, the function value $f(\lim_{n\to\infty} a_n)$ is well defined since we assume that $f : \mathbb{Q} \to \mathbb{Q}$.

We see that it is sufficient that $f(x)$ is Lipschitz continuous on an interval I containing all the elements a_n as well as $\lim_{n\to\infty} a_n$. We have thus proved the following fundamental result.

Theorem 13.1 *Let $\{a_n\}$ be a sequence with rational limit $\lim_{n\to\infty} a_n$. Let $f : I \to \mathbb{Q}$ be a Lipschitz continuous function, and assume that $a_n \in I$ for all n and $\lim_{n\to\infty} a_n \in I$. Then,*

$$\lim_{n\to\infty} f(a_n) = f\left(\lim_{n\to\infty} a_n\right). \tag{13.8}$$

Note that choosing I to be a closed interval guarantees that $\lim_{n\to\infty} a_n \in I$ if $a_n \in I$ for all n.

We now look at some examples.

EXAMPLE 13.4. In the growth of bacteria model of Chapter Rational Numbers, we need to compute

$$\lim_{n\to\infty} P_n = \lim_{n\to\infty} \frac{1}{\frac{1}{2^n}Q_0 + \frac{1}{K}\left(1 - \frac{1}{2^n}\right)}.$$

The sequence $\{P_n\}$ is obtained by applying the function

$$f(x) = \frac{1}{Q_0 x + \frac{1}{K}(1-x)}$$

to the terms in the sequence $\{\frac{1}{2^n}\}$. Since $\lim_{n\to\infty} 1/2^n = 0$, we have $\lim_{n\to\infty} P_n = f(0) = K$, since f is Lipschitz continuous on for example $[0, 1/2]$. The Lipschitz continuity follows from the fact that $f(x)$ is the composition of the function $f_2(x) = Q_0 x + \frac{1}{K}(1-x)$ and the function $f_2(y) = 1/y$.

EXAMPLE 13.5. The function $f(x) = x^2$ is Lipschitz continuous on bounded intervals. We conclude that if $\{a_n\}_{n=1}^\infty$ converges to a rational limit A, then

$$\lim_{n\to\infty} \big(a_n\big)^2 = A^2.$$

In the next chapter, we will be interested in computing $\lim_{n\to\infty} \big(a_n\big)^2$ for a certain sequence $\{a_n\}_{n=1}^\infty$ arising in connection with the Muddy Yard model, which will bring a surprise. Can you guess what it is?

EXAMPLE 13.6. By Theorem 13.1, with appropriate choices (which?) of the function $f(x)$:

$$\lim_{n\to\infty} \left(\frac{3+\frac{1}{n}}{4+\frac{2}{n}}\right)^9 = \left(\lim_{n\to\infty} \frac{3+\frac{1}{n}}{4+\frac{2}{n}}\right)^9 = \left(\frac{\lim_{n\to\infty}(3+\frac{1}{n})}{\lim_{n\to\infty}(4+\frac{2}{n})}\right)^9 = \left(\frac{3}{4}\right)^9.$$

EXAMPLE 13.7. By Theorem 13.1,

$$\lim_{n\to\infty} (2^{-n})^7 + 14(2^{-n})^4 - 3(2^{-n}) + 2$$
$$= 0^7 + 14 \times 0^4 - 3 \times 0 + 2 = 2.$$

13.8 Generalization to functions of two variables

We recall that a function $f : I \times J \to \mathbb{Q}$ of two rational variables, where I and J are closed intervals of $\mathbb{Q}$, is said to be Lipschitz continuous if there is constant L such that

$$|f(x_1, x_2) - f(\bar{x}_1, \bar{x}_2)| \leq L(|x_1 - \bar{x}_1| + |x_2 - \bar{x}_2|)$$

for $x_1, \bar{x}_1 \in I$ and $x_2, \bar{x}_2 \in J$.

Let now $\{a_n\}$ and $\{b_n\}$ be two converging sequences of rational numbers with $a_n \in I$ and $b_n \in J$. Then

$$f(\lim_{n\to\infty} a_n, \lim_{n\to\infty} b_n)) = \lim_{n\to\infty} f(a_n, b_n). \tag{13.9}$$

The proof is immediate:

$$|f(\lim_{n\to\infty} a_n, \lim_{n\to\infty} b_n)) - f(a_m, b_m)| \leq L(|\lim_{n\to\infty} a_n - a_m| + |\lim_{n\to\infty} b_n - b_m|) \tag{13.10}$$

where the right hand side can be made arbitrarily small by choosing m large enough.

We give a first application of this result with $f(x_1, x_2) = x_1 + x_2$, which is Lipschitz continuous on $\mathbb{Q} \times \mathbb{Q}$ with Lipschitz constant $L = 1$. We conclude from (13.9) the natural formula

$$\lim_{n\to\infty}(a_n + b_n) = \lim_{n\to\infty} a_n + \lim_{n\to\infty} b_n \tag{13.11}$$

stating that the limit of the sum is the sum of the limits.

Similarly we have, of course, using the function $f(x_1, x_2) = x_1 - x_2$,

$$\lim_{n\to\infty}(a_n - b_n) = \lim_{n\to\infty} a_n - \lim_{n\to\infty} b_n.$$

As a special case choosing $a_n = a$ for all n,

$$\lim_{n\to\infty}(a + b_n) = a + \lim_{n\to\infty} b_n.$$

Next, we consider the function $f(x_1, x_2) = x_1 x_2$, which is Lipschitz continuous on $I \times J$, if I and J are closed bounded intervals of $\mathbb{Q}$. Using (13.9) we find that if $\{a_n\}$ and $\{b_n\}$ are two converging sequences of rational numbers, then

$$\lim_{n\to\infty}(a_n \times b_n) = \lim_{n\to\infty} a_n \times \lim_{n\to\infty} b_n,$$

stating that the limit of the products is the product of the limits.

As a special case choosing $a_n = a$ for all n, we have

$$\lim_{n\to\infty}(a \times b_n) = a \lim_{n\to\infty} b_n.$$

We now consider the function $f(x_1, x_2) = x_1/x_2$, which is Lipschitz continuous on $I \times J$, if I and J are closed intervals of $\mathbb{Q}$ with J not including 0. If $b_n \in J$ for all n and $\lim_{n\to\infty} b_n \neq 0$, then

$$\lim_{n\to\infty}(a_n/b_n) = \frac{\lim_{n\to\infty} a_n}{\lim_{n\to\infty} b_n},$$

stating that the limit of the quotient is the quotient of the limits if the limit of the denominator is not zero.

13.9 Computing limits

We now apply the above rules to compute some limits.

EXAMPLE 13.8. Consider $\{2 + 3n^{-4} + (-1)^n n^{-1}\}_{n=1}^{\infty}$.

$$\begin{aligned}
&\lim_{n\to\infty}(2 + 3n^{-4} + (-1)^n n^{-1}) \\
&\qquad = \lim_{n\to\infty} 2 + 3\lim_{n\to\infty} n^{-4} + \lim_{n\to\infty}(-1)^n n^{-1} \\
&\qquad = 2 + 3 \times 0 + 0 = 2.
\end{aligned}$$

To do this example, we use (13.11) and the fact that

$$\lim_{n\to\infty} n^{-p} = \big(\lim_{n\to\infty} n^{-1}\big)^p = 0^p = 0$$

for any natural number p.

Another useful fact is

$$\lim_{n\to\infty} r^n = \begin{cases} 0 & \text{if } |r| < 1, \\ 1 & \text{if } r = 1, \\ \text{diverges to } \infty & \text{if } r > 1, \\ \text{diverges} & \text{otherwise.} \end{cases}$$

We showed the case when $r = 1/2$ in Example 13.4 and you will show the general result later as an exercise.

EXAMPLE 13.9. Using 13.3, we can solve for the limiting behavior of the population of bacteria described in Example 13.4. We have

$$\begin{aligned} \lim_{n\to\infty} P_n &= \frac{1}{\lim_{n\to\infty} \frac{1}{2^n} Q_0 + \lim_{n\to\infty} \frac{1}{K}\left(1 - \frac{1}{2^n}\right)} \\ &= \frac{1}{0 + \frac{1}{K}(1-0)} = K. \end{aligned}$$

In words, the population of the bacteria growing under the limited resources as modeled by the Verhulst model tends to a constant population.

EXAMPLE 13.10. Consider

$$\left\{ 4\,\frac{1+n^{-3}}{3+n^{-2}} \right\}_{n=1}^{\infty}.$$

We compute the limit using the different rules:

$$\begin{aligned} \lim_{n\to\infty} 4\,\frac{1+n^{-3}}{3+n^{-2}} &= 4\,\frac{\lim_{n\to\infty}\big(1+n^{-3}\big)}{\lim_{n\to\infty}\big(3+n^{-2}\big)} \\ &= 4\,\frac{1+\lim_{n\to\infty} n^{-3}}{3+\lim_{n\to\infty} n^{-2}} \\ &= 4\,\frac{1+0}{3+0} = \frac{4}{3}. \end{aligned}$$

EXAMPLE 13.11. Consider

$$\left\{ \frac{6n^2+2}{4n^2-n+1000} \right\}_{n=1}^{\infty}.$$

Before computing the limit, think about what is going on as n becomes large. In the numerator, $6n^2$ is much larger than 2 when n is large and likewise in the denominator, $4n^2$ becomes much larger than $-n+1000$ in size when n is large. So we might guess that for n large,

$$\frac{6n^2+2}{4n^2-n+1000} \approx \frac{6n^2}{4n^2} = \frac{6}{4}.$$

This would be a good guess for the limit. To see that this is true, we use a trick to put the sequence in a better form to compute the limit:

$$\begin{aligned}
\lim_{n\to\infty} \frac{6n^2+2}{4n^2-n+1000} &= \lim_{n\to\infty} \frac{(6n^2+2)n^{-2}}{(4n^2-n+1000)n^{-2}} \\
&= \lim_{n\to\infty} \frac{6+2n^{-2}}{4-n^{-1}+1000n^{-2}} \\
&= \frac{6}{4}
\end{aligned}$$

where we finished the computation as in the previous example.

The trick of multiplying top and bottom of a ratio by a power can also be used to figure out when a sequence converges to zero or diverges to infinity.

Example 13.12.

$$\begin{aligned}
\lim_{n\to\infty} \frac{n^3-20n^2+1}{n^8+2n} &= \lim_{n\to\infty} \frac{(n^3-20n^2+1)n^{-3}}{(n^8+2n)n^{-3}} \\
&= \lim_{n\to\infty} \frac{1-20n^{-1}+n^{-3}}{n^5+2n^{-2}}.
\end{aligned}$$

From this we see that the numerator converges to 1 while the denominator increases without bound. Therefore

$$\lim_{n\to\infty} \frac{n^3-20n^2+1}{n^8+2n} = 0.$$

Example 13.13.

$$\begin{aligned}
\lim_{n\to\infty} \frac{-n^6+n+10}{80n^4+7} &= \lim_{n\to\infty} \frac{(-n^6+n+10)n^{-4}}{(80n^4+7)n^{-4}} \\
&= \lim_{n\to\infty} \frac{-n^2+n^{-3}+10n^{-4}}{80+7n^{-4}}.
\end{aligned}$$

From this we see that the numerator grows in the negative direction without bound while the denominator tends towards 80. Therefore

$$\left\{\frac{-n^6+n+10}{80n^4+7}\right\}_{n=1}^{\infty} \quad \text{diverges to } -\infty.$$

13.10 Computer representation of rational numbers

The decimal expansion $\pm p_m p_{m-1} \cdots p_1 p_0 . q_1 q_2 \cdots q_n$ uses the base 10 and consequently each of the digits p_i and q_j may take on one of the 10 values $0, 1, 2, ...9$. Of course, it is possible to use bases other than ten. For example, the Babylonians used the base sixty and thus their digits range between 0 and 59. The computer operates with the base 2 and the two digits 0 and 1. A base 2 number has the form

$$\pm p_m 2^m + p_{m-1} 2^{m-1} + ... + p_2 2^2 + p_1 2^1 + p_0 2^0 + q_1 2^{-1} + q_2 2^{-2} \\ + ... + q_{n-1} 2^{-(n-1)} + q_n 2^{-n},$$

which we again may write in short hand

$$\pm p_m p_{m-1} ... p_1 p_0 . q_1 q_2 .... q_n = p_m p_{m-1} ... p_1 p_0 + 0.q_1 q_2 .... q_n$$

where again n and m are natural numbers, and now each p_i and q_j take the value 0 or 1. For example, in the base two

$$11.101 = 1 \cdot 2^1 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 1 \cdot 2^{-3}.$$

In the floating point arithmetic of a computer using the standard 32 bits, numbers are represented in the form

$$\pm r 2^N,$$

where $1 \leq r \leq 2$ is the *mantissa* and the *exponent* N is an integer. Out of the 32 bits, 23 bits are used to store the mantissa, 8 bits are used to store the exponent and finally one bit is used to store the sign. Since $2^{10} \approx 10^{-3}$ this gives 6 to 7 decimal digits for the mantissa while the exponent N may range from -126 to 127, implying that the absolute value of numbers stored on the computer may range from approximately 10^{-40} to 10^{40}. Numbers outside these ranges cannot be stored by a computer using 32 bits. Some languages permit the use of double precision variables using 64 bits for storage with 11 bits used to store the exponent, giving a range of $-1022 \leq n \leq 1023$, 52 bits used to store the the mantissa, giving about 15 decimal places.

We point out that the finite storage capability of a computer has two effects when storing rational numbers. The first effect is similar to the effect of finite storage on integers, namely only rational numbers within a finite range can be stored. The second effect is more subtle but actually has more serious consequences. This is the fact that numbers are stored only up to a specified number of digits. Any rational number that requires more than the finite number of digits in its decimal expansions, which included all rational numbers with infinite periodic expansions for example, are therefore stored on a computer with an error. So for example $2/11$ is stored as .1818181 or .1818182 depending on whether the computer rounds or not.

But this is not the end of the story. Introduction of an error in the 7'th or 15'th digit would not be so serious except for the fact that such *round-off* errors accumulate when arithmetic operations are performed. In other words, if we add two numbers with a small error, the result may have a larger error being the sum of the individual errors (unless the errors have opposite sign or even cancel).

We give below in Chapter Series an example showing a startling consequence of working with finite decimal representations with round off errors.

13.11 Sonya Kovalevskaya: The first woman on a chair in mathematics

Sonya Kovalevskaya (1850-91) was a student of Weierstrass and as the first woman ever got a position 1889 as Professor in Mathematics at the University of Stockholm, see Fig. 13.4. Her mentor was Gösta Mittag-Leffler (1846-1927), famous Swedish mathematician and founder of the prestigous journal Acta Mathematica, see Fig. 22.4.

Kovalevskaya was 1886 awarded the 5,000 francs Prix Bordin for her paper *Mèmoire sur un cas particulier du problème de le rotation d'un corps pesant autour d'un point fixe, ou l'intgration s'effectue l'aide des fonctions ultraelliptiques du temps.* At the height of her career, Kovalevskaya died of influenza complicated by pneumonia, only 41 years old.

Chapter 13 Problems

13.1. Plot the functions; (a) 2^{-n}, (b) 5^{-n}, and (c) 10^{-n}; defined on the natural numbers n. Compare the plots.

13.2. Plot the function $f(n) = \frac{10}{9}(1 - 10^{-n-1})$ defined on the natural numbers.

13.3. Write the following sequences using the index notation:

(a) $\{1, 3, 9, 27, \cdots\}$ (b) $\{16, 64, 256, \cdots\}$

(c) $\{1, -1, 1, -1, 1, \cdots\}$ (d) $\{4, 7, 10, 13, \cdots\}$

(e) $\{2, 5, 8, 11, \cdots\}$ (f) $\{125, 25, 5, 1, \frac{1}{5}, \frac{1}{25}, \frac{1}{125}, \cdots\}$.

13.4. Show the following limits hold using the formal definition of the limit:

$$\text{(a)} \lim_{n\to\infty} \frac{8}{3n+1} = 0 \quad \text{(b)} \lim_{n\to\infty} \frac{4n+3}{7n-1} = \frac{4}{7} \quad \text{(c)} \lim_{n\to\infty} \frac{n^2}{n^2+1} = 1.$$

FIGURE 13.4. Sonya Kovalevskaya, first woman as a Professor of Mathematics: "I began to feel an attraction for my mathematics so intense that I started to neglect my other studies" (at age 11).

13.5. Show that $\lim_{n\to\infty} r^n = 0$ for any r with $|r| \le 1/2$.

13.6. One of the classic paradoxes posed by the Greek philosophers can be solved using the geometric series. Suppose you are in Paulding county on your bike, 32 miles from home. You break a spoke, you have no more food and you drank the last of your water, you forgot to bring money and it starts to rain. While riding home, as wont to do, you begin to think about how far you have to ride. Then you have a depressing thought: you can never get home! You think to yourself: first I have to ride 16 miles, then 8 miles after that, then 4 miles, then 2, then 1, then 1/2, then 1/4, and so on. Apparently you always have a little way to go, no matter how close you are, and you have to add up an infinite number of distances to get anywhere! The Greek philosophers did not understand how to interpret a limit of a sequence, so this caused them a great deal of trouble. Explain why there is no paradox involved here using the sum of a geometric series.

13.7. Show the following hold using the formal definition for divergence to infinity:

$$\text{(a)}\ \lim_{n\to\infty} -4n+1 = -\infty \qquad \text{(b)}\ \lim_{n\to\infty} n^3+n^2 = \infty.$$

13.8. Show that $\lim_{n\to\infty} r^n = \infty$ for any r with $|r| \ge 2$.

13.9. Find the values of

(a) $1 - .5 + .25 - .125 + \cdots$

(b) $3 + \dfrac{3}{4} + \dfrac{3}{16} + \cdots$

(c) $5^{-2} + 5^{-3} + 5^{-4} + \cdots$

13.10. Find formulas for the sums of the following series by using the formula for the sum of the geometric series assuming $|r| < 1$:

(a) $1 + r^2 + r^4 + \cdots$

(b) $1 - r + r^2 - r^3 + r^4 - r^5 + \cdots$

13.11. Determine the number of different sequences there are in the following list and identify the sequences that are equal.

$$\text{(a)}\ \left\{\frac{4^{n/2}}{4+(-1)^n}\right\}_{n=1}^{\infty} \qquad \text{(b)}\ \left\{\frac{2^n}{4+(-1)^n}\right\}_{n=1}^{\infty}$$

$$\text{(c)}\ \left\{\frac{2^{\,\mathrm{car}}}{4+(-1)^{\,\mathrm{car}}}\right\}_{\mathrm{car}=1}^{\infty} \qquad \text{(d)}\ \left\{\frac{2^{n-1}}{4+(-1)^{n-1}}\right\}_{n=2}^{\infty}$$

$$\text{(e)}\ \left\{\frac{2^{n+2}}{4+(-1)^{n+2}}\right\}_{n=0}^{\infty} \qquad \text{(f)}\ \left\{8\frac{2^n}{4+(-1)^{n+3}}\right\}_{n=-2}^{\infty}.$$

13.12. Rewrite the sequence $\left\{\dfrac{2+n^2}{9^n}\right\}_{n=1}^{\infty}$ so that: (a) the index n runs from -4 to ∞, (b) the index n runs from 3 to ∞, (c) the index n runs from 2 to $-\infty$.

13.13. Show that (7.14) holds by considering the different cases: $a < 0$, $b < 0$, $a < 0$, $b > 0$, $a > 0$, $b < 0$, $a > 0$, $b > 0$. Show that (13.7) holds using (7.14) and the fact that $a - c + c - b = a - b$.

13.14. Suppose that $\{a_n\}_{n=1}^{\infty}$ converges to A and $\{b_n\}_{n=1}^{\infty}$ converges to B. Show that $\{a_n - b_n\}_{n=1}^{\infty}$ converges to $A - B$.

13.15. *(Harder)* Suppose that $\{a_n\}_{n=1}^{\infty}$ converges to A and $\{b_n\}_{n=1}^{\infty}$ converges to B. Show that if $b_n \neq 0$ for all n and $B \neq 0$, then $\{a_n/b_n\}_{n=1}^{\infty}$ converges to A/B. Hint: write

$$\frac{a_n}{b_n} - \frac{A}{B} = \frac{a_n}{b_n} + \frac{a_n}{B} - \frac{a_n}{B} - \frac{A}{B}$$

and the fact that for n large enough, $|b_n| \geq B/2$. Be sure to say why the last fact is true!

13.16. Compute the limits of the sequences $\{a_n\}_{n=1}^{\infty}$ with the indicated terms or show they diverge.

(a) $a_n = 1 + \dfrac{7}{n}$ (b) $a_n = 4n^2 - 6n$

(c) $a_n = \dfrac{(-1)^n}{n^2}$ (d) $a_n = \dfrac{2n^2 + 9n + 3}{6n^2 + 2}$

(e) $a_n = \dfrac{(-1)^n n^2}{7n^2 + 1}$ (f) $a_n = \left(\dfrac{2}{3}\right)^n + 2$

(g) $a_n = \dfrac{(n-1)^2 - (n+1)^2}{n}$ (h) $a_n = \dfrac{1 - 5n^8}{4 + 51n^3 + 8n^8}$

(i) $a_n = \dfrac{2n^3 + n + 1}{6n^2 - 5}$ (j) $a_n = \dfrac{\left(\frac{7}{8}\right)^n - 1}{\left(\frac{7}{8}\right)^n + 1}$.

13.17. Compute the following limits

(a) $\displaystyle\lim_{n\to\infty} \left(\frac{n+3}{2n+8}\right)^{37}$ (b) $\displaystyle\lim_{n\to\infty} \left(\frac{31}{n^2} + \frac{2}{n} + 7\right)^4$

(c) $\displaystyle\lim_{n\to\infty} \frac{1}{\left(2 + \frac{1}{n}\right)^8}$ (d) $\displaystyle\lim_{n\to\infty} \left(\left(\left(\left(1 + \frac{2}{n}\right)^2\right)^3\right)^4\right)^5$.

13.18. Rewrite the following sequences as a function applied to another sequence three different ways:

(a) $\displaystyle\left\{\left(\frac{n^2+2}{n^2+1}\right)^3\right\}_{n=1}^{\infty}$. (b) $\displaystyle\left\{\left(n^2\right)^4 + \left(n^2\right)^2 + 1\right\}_{n=1}^{\infty}$

13.19. Show that the infinite decimal expansion 0.9999.... is equal to 1. In other words, show that

$$\lim_{n\to\infty} 0.99\cdots 99_n = 1,$$

where $0.99\cdots 99_n$ contains n decimals all equal to 9.

13.20. Determine the number of digits used to store rational numbers in the programming language that you use and whether the language truncates or rounds.

13.21. The *machine number* u is the smallest positive number u stored in a computer that satisfies $1+u > 1$. Note that u is not zero! For example in a single precision language $1+.00000000001 = 1$, explain why. Write a little program that computes the u for your computer and programming language. Hint: $1 + .5 > 1$ in any programming language. Also $1 + .25 > 1$. Continue...

14
The squareroot of two

He is unworthy of the name man who is ignorant of the fact that the diagonal of a square is incommensurable with its side. (Plato)

Just as the introduction of the irrational number is a convenient myth which simplifies the laws of arithmetics...so physical objects are postulated entities which round out and simplify our account of the flux of existence... The conceptual scheme of physical objects is likewise a convenient myth, simpler than the literal truth and yet containing that literal truth as a scattered part. (Quine)

14.1 Introduction

We met the equation $x^2 = 2$ in the context of the Muddy Yard model, trying to determine the length of the diagonal of a square with side length 1. We have learned in school that the (positive) solution of the equation $x^2 = 2$ is $x = \sqrt{2}$. But, honestly speaking, what *is* in fact $\sqrt{2}$? To simply say that it is the solution of the equation $x^2 = 2$, or "that number which when squared is equal to 2", leads to circular reasoning, and would not help much when trying to by a pipe of length $\sqrt{2}$.

We then may recall again from school that $\sqrt{2} \approx 1.41$, but computing $1.41^2 = 1.9881$, we see that $\sqrt{2}$ is not exactly equal to 1.41. A better guess is 1.414, but then we get $1.414^2 = 1.999386$. We use *MAPLE*© to compute the decimal expansion of $\sqrt{2}$ to 415 places:

$$
\begin{aligned}
x = 1.&4142135623730950488016887242096980785696718753\\
&7694807317667973799073247846210703885038753432\\
&7641572735013846230912297024924836055850737212\\
&6441214970999358314132226659275055927557999505\\
&0115278206057147010955997160597027453459686201\\
&4728517418640889198609552329230484308714321450\\
&8397626036279952514079896872533965463318088296\\
&4062061525835239505474575028775996172983557522\\
&0337531857011354374603408498847160386899970699
\end{aligned}
$$

Computing x^2 again using *MAPLE*© , we find that

$$
\begin{aligned}
x^2 = 1.&999999999999999999999999999999999999999999999\\
&999999999999999999999999999999999999999999999\\
&999999999999999999999999999999999999999999999\\
&999999999999999999999999999999999999999999999\\
&999999999999999999999999999999999999999999999\\
&999999999999999999999999999999999999999999999\\
&999999999999999999999999999999999999999999999\\
&999999999999999999999999999999999999999999999\\
&999999999999999999999999999999999999999999999\\
&999999999986381037002790393547544921481567520\\
&719364336722392248627179189098787015809960232\\
&640597261312640760405691299950309295747831888\\
&596950070887405605833650165227157380944559332\\
&069004581726422217393596953324251515876023360\\
&427299488914180359897103820495618481233332162\\
&516016097283137123064499497943653479698629776\\
&683334066577024031851330600242723212517527304\\
&354776748660808998780793579777475964587708250\\
&3170068870585486010
\end{aligned}
$$

The number $x = 1.4142\cdots 699$ satisfies the equation $x^2 = 2$ to a high degree of precision but not exactly. In fact, it turns out that no matter how many digits we take in a guess of $\sqrt{2}$ with a finite decimal expansion, we never get a number which squared gives exactly 2. So it seems that we have not yet really caught the exact value of $\sqrt{2}$. So what is it?

To get a clue, we may try to examine the decimal expansion of $\sqrt{2}$, but we will not find any pattern. In particular, the first 415 places show no periodic pattern.

14.2 $\sqrt{2}$ is not a rational number!

In this section, we show that $\sqrt{2}$ cannot be a rational number of the form p/q with p and q natural numbers, and thus the decimal expansion of $\sqrt{2}$ cannot be periodic. In the proof we use the fact that a natural number can be uniquely factored into prime factors. We showed this in chapter Natural Numbers and Integers. One consequence of the factorization into prime numbers is the following fact: Suppose that we know that 2 is a factor of n. If $n = pq$ is a factorization of n into integers p and q, if follows that at least one of the factors p and q must have a factor of 2.

We argue by contradiction. Thus we shall show that assuming that $\sqrt{2}$ is rational leads to a contradiction, and thus $\sqrt{2}$ cannot be rational. We thus assume that $\sqrt{2} = p/q$, where all common factors in the natural numbers p and q have been divided out. For example if p and q both have the factor 3, then we replace p by $p/3$ and q by $q/3$, which does not change the quotient p/q. We write this as $\sqrt{2}q = p$ where p and q have no common factors, or squaring both sides, $2q^2 = p^2$. Since the left hand side contains the factor 2, the right hand side p^2 must contain the factor 2, which means that p must contain the factor 2. Thus we can write $p = 2 \times \bar{p}$ with $\bar{p}$ a natural number. We conclude that $2q^2 = 4 \times \bar{p}^2$, that is $q^2 = 2 \times \bar{p}^2$. But the same argument implies that q must also contain a factor of 2. Thus both p and q contain the factor 2 which contradicts the original assumption that p and q had no common factors. Assuming $\sqrt{2}$ to be rational number thus leads to a contradiction and therefore $\sqrt{2}$ cannot be a rational number.

The argument just given was known to the Pythagoreans, who thus knew that $\sqrt{2}$ is not a rational number. This knowledge caused a lot of trouble. On one hand, $\sqrt{2}$ represents the diagonal of a square of side one, so it seemed that $\sqrt{2}$ had to exist. On the other hand, the Pythagorean school of philosophy was based on the principle that everything could be described in terms of natural numbers. The discovery that $\sqrt{2}$ was not a rational number, that is that $\sqrt{2}$ could not be viewed as a pair of natural numbers, came as a shock! Legend says that the person who discovered the proof was punished by the gods for revealing an imperfection in the universe. The Pythagoreans tried to keep the discovery secret by teaching it only to a select few, but eventually the discovery was revealed and after that the Pythagorean school quickly fell apart. At the same time, the Euclidean school, which was based on geometry instead of numbers, became more influential. Considered from the point of view of geometry, the difficulty with $\sqrt{2}$ seems to "disappear", because no one would question that a square of side length 1 will have a diagonal of a certain length, and we could then simply define $\sqrt{2}$ to be that length. The Euclidean geometric school took over and ruled all through the Dark Ages until the time of Descartes in the 17th century who resurrected the Pythagorean school based on numbers, in the form of analytical geometry. Since the digital computer of today is based on natural numbers, or rather sequences of $0s$ and $1s$, we may say that Pythagoras ideas are very much alive today: everything can be described in terms of natural numbers. Other Pythagorean dogmas like "never eat beans" and "never pick up anything that has fallen down" have not survived equally well.

14.3 Computing $\sqrt{2}$ by the Bisection algorithm

We now present an algorithm for computing a sequence of rational numbers that satisfy the equation $x^2 = 2$ more and more accurately. That is, we construct a sequence of rational number approximations of a solution of the equation

$$f(x) = 0 \tag{14.1}$$

with $f(x) = x^2 - 2$. The algorithm uses a trial and error strategy that checks whether a given number r satisfies $f(r) < 0$ or $f(r) > 0$, i.e. if $r^2 < 2$ or $r^2 > 2$. All of the numbers r constructed during this process are rational, so none of them can ever actually equal $\sqrt{2}$.

We begin by noting that $f(1) < 0$ since $1^2 < 2$ and $f(2) > 0$ since $2^2 > 2$. Now since $0 < x < y$ means that $x^2 < xy < y^2$, we know that $f(x) < 0$ for all $0 < x \leq 1$ and $f(x) > 0$ for all $x \geq 2$. So any solution of (14.1) must lie between 1 and 2. Hence we choose a point between 1 and 2 and check the sign of f at that point. For the sake of symmetry, we choose the halfway point $1.5 = (1+2)/2$ of 1 and 2. We find that $f(1.5) > 0$. Remembering that $f(1) < 0$, we conclude that a (positive) solution of (14.1) must lie between 1 and 1.5.

We continue, next checking the mean value 1.25 of 1 and 1.5 to find that $f(1.25) < 0$. This means that a solution of (14.1) must lie between 1.25 and 1.5. Next we choose the point halfway between these two, 1.375, and find that $f(1.375) < 0$, implying that any solution of (14.1) lies between 1.375 and 1.5. We can continue to search in this way as long as we like, each time determining two rational numbers that "trap" any solution of (14.1). This process is called the *Bisection Algorithm*.

1. Choose the initial values x_0 and X_0 so that $f(x_0) < 0$ and $f(X_0) > 0$. Set $i = 1$.

2. Given two rational numbers x_{i-1} and X_{i-1} with the property that $f(x_{i-1}) < 0$ and $f(X_{i-1}) > 0$, set $\bar{x}_i = (x_{i-1} + X_{i-1})/2$.

- If $f(\bar{x}_i) = 0$, then stop.
- If $f(\bar{x}_i) < 0$, then set $x_i = \bar{x}_i$ and $X_i = X_{i-1}$.
- If $f(\bar{x}_i) > 0$, then set $x_i = x_{i-1}$ and $X_i = \bar{x}_i$.

3. Increase i by 1 and go back to step 2.

We list the output for 20 steps from a *MATLAB*© *m*-file implementing this algorithm in Fig. 14.1 with $x_0 = 1$ and $X_0 = 2$.

i	x_i	X_i
0	1.00000000000000	2.00000000000000
1	1.00000000000000	1.50000000000000
2	1.25000000000000	1.50000000000000
3	1.37500000000000	1.50000000000000
4	1.37500000000000	1.43750000000000
5	1.40625000000000	1.43750000000000
6	1.40625000000000	1.42187500000000
7	1.41406250000000	1.42187500000000
8	1.41406250000000	1.41796875000000
9	1.41406250000000	1.41601562500000
10	1.41406250000000	1.41503906250000
11	1.41406250000000	1.41455078125000
12	1.41406250000000	1.41430664062500
13	1.41418457031250	1.41430664062500
14	1.41418457031250	1.41424560546875
15	1.41418457031250	1.41421508789062
16	1.41419982910156	1.41421508789062
17	1.41420745849609	1.41421508789062
18	1.41421127319336	1.41421508789062
19	1.41421318054199	1.41421508789062
20	1.41421318054199	1.41421413421631

FIGURE 14.1. 20 steps of the Bisection algorithm.

14.4 The Bisection algorithm converges!

By continuing the Bisection Algorithm without stopping, we generate two sequences of rational numbers $\{x_i\}_{i=0}^{\infty}$ and $\{X_i\}_{i=0}^{\infty}$. By construction,

$$\begin{gathered} x_0 \leq x_1 \leq x_2 \leq \cdots \quad \text{and} \quad X_0 \geq X_1 \geq X_2 \geq \cdots \\ x_i < X_j \quad \text{for all } i, j = 0, 1, 2, \ldots \end{gathered}$$

In other words, the terms x_i either increase or stay constant while the X_i always decrease or remain constant as i increases, and any x_i is smaller than any X_j. Moreover, the choice of the midpoint means that the distance between X_i and x_i is always strictly decreasing as i increases. In fact,

$$0 \le X_i - x_i \le 2^{-i} \quad \text{for } i = 0, 1, 2, \cdots, \tag{14.2}$$

i.e. the difference between the value x_i for which $f(x_i) < 0$ and the value X_i for which $f(X_i) > 0$ is halved for each step increase i by 1. This means that as i increases, more and more digits in the decimal expansions of x_i and X_i agree. Since $2^{-10} \approx 10^{-3}$, we gain approximately 3 decimal places for every 10 successive steps of the bisection algorithm. We can see this in Fig. 14.1.

The estimate (14.2) on the difference of $X_i - x_i$ also implies that the terms in the sequence $\{x_i\}_{i=0}^{\infty}$ become closer as the index increase. This follows because $x_i \le x_j < X_j \le X_i$ if $j > i$ so (14.2) implies

$$|x_i - x_j| \le |x_i - X_i| \le 2^{-i} \quad \text{if } j \ge i.$$

that is

$$|x_i - x_j| \le 2^{-i} \quad \text{if } j \ge i. \tag{14.3}$$

We illustrate in Fig. 14.2. In particular, this means that when $2^{-i} \le$

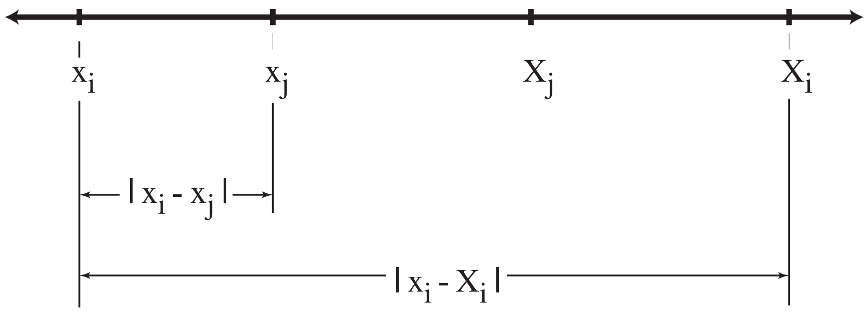


FIGURE 14.2. $|x_i - x_j| \le |X_i - x_i|$.

10^{-N-1}, the first N decimals of x_j are the same as the first N decimals in x_i for any $j \ge i$.

In other words, as we compute more and more numbers x_i, more and more leading decimals of the numbers x_i agree. We conclude that the sequence $\{x_i\}_{i=0}^{\infty}$ determines a specific (infinite) decimal expansion. To get the first N digits of this expansion, we simply take the first N digits of any number x_j in the sequence with $2^{-j} \le 10^{-N-1}$. By the inequality (14.3), all such x_j agree in the first N digits.

If this infinite decimal expansion was the decimal expansion of a rational number $\bar{x}$, then we would of course have

$$\bar{x} = \lim_{i\to\infty} x_i.$$

However, we showed above that the decimal expansion defined by the sequence $\{x_i\}_{i=0}^{\infty}$ cannot be periodic. So there is no rational number $\bar{x}$ that can be the limit of the sequence $\{x_i\}$.

We have now come to the point where the Pythagoreans got stuck 2.500 years ago. The sequence $\{x_i\}$ "tries to converge" to a limit, but the limit is not a number of the type we already know, that is a rational number. To avoid the fate of the Pythagoreans, we have to find a way out of this dilemma. The limit appears to be a number of a new kind and thus it appears that we have to somehow extend the rational numbers. The extension will be accomplished by viewing any infinite decimal expansion, periodic or not, as some kind of number, more precisely as a *real number*. In this way, we will clearly get an extension of the set of rational numbers since the rational numbers correspond to periodic decimal expansions. We will refer to non-periodic decimal expansions as *irrational numbers*.

For the extension from rational to real numbers to make sense, we must show that we can compute with irrational numbers in pretty much the same way as with rational numbers. We shall see this is indeed possible and we shall see that the basic idea when computing with irrational numbers is the natural one: compute with truncated decimal expansions! We give the details in the next chapter devoted to a study of real numbers.

Let us now summarize and see where we stand: the Bisection algorithm applied to the equation $x^2 - 2 = 0$ generates a sequence $\{x_i\}_{i=1}^{\infty}$ satisfying

$$|x_i - x_j| \leq 2^{-i} \quad \text{if } j \geq i. \tag{14.4}$$

The sequence $\{x_i\}_{i=1}^{\infty}$ defines an infinite non-periodic decimal expansion, which we will view as an irrational number. We will give this irrational number the *name* $\sqrt{2}$. We thus use $\sqrt{2}$ as a *symbol* to denote a certain infinite decimal expansion determined by the Bisection algorithm applied to the equation $x^2 - 2 = 0$.

We now need to specify how to compute with irrational numbers. Once we have done this, it remains to show that the particular irrational number named $\sqrt{2}$ constructed above indeed does satisfy the equation $x^2 = 2$. That is after defining multiplication of irrational numbers like $\sqrt{2}$, we need to show that

$$\sqrt{2}\sqrt{2} = 2. \tag{14.5}$$

Note that this equality does not follow directly by definition, as it would if we had defined $\sqrt{2}$ as "that thing" which multiplied with itself equals 2 (which doesn't make sense since we don't know that "that thing" exists). Instead, we have now defined $\sqrt{2}$ as the infinite decimal expansion defined by the Bisection algorithm applied to $x^2-2=0$, and it is a non-trivial step to first define what we mean by multiplying $\sqrt{2}$ by $\sqrt{2}$, and then to show that indeed $\sqrt{2}\sqrt{2}=2$. This is what the Pythagoreans could not manage to do, which had devastating effects on their society.

We return to verifying (14.5) after showing in the next chapter how to compute with real numbers, so that in particular we know how to multiply the irrational number $\sqrt{2}$ with itself!

14.5 First encounters with Cauchy sequences

We recall that the sequence $\{x_i\}$ defined by the Bisection algorithm for solving the equation $x^2=2$, satisfies

$$|x_i - x_j| \leq 2^{-i} \quad \text{if } j \geq i, \tag{14.6}$$

from which we concluded that the sequence $\{x_i\}_{i=1}^{\infty}$ specifies a certain infinite decimal expansion. To get the first N decimals of the expansion we take the first N decimals of any number x_j in the sequence with $2^{-j} \leq 10^{-N-1}$. Any two such x_j will agree to N decimals in the sense that their difference is a most 1 in decimal place $N+1$.

The sequence $\{x_i\}$ satisfying (14.6) is an example of a Cauchy sequence of rational numbers. More generally, a sequence $\{y_i\}$ of rational numbers is said to be a *Cauchy sequence* if for any $\epsilon > 0$ there is a natural number N such that

$$|y_i - y_j| \leq \epsilon \quad \text{if } i, j \geq N.$$

To show that the sequence $\{x_i\}$ satisfying (14.6) is indeed a Cauchy sequence, we first choose $\epsilon > 0$ and then we choose N so that $2^{-N} \leq \epsilon$.

As a basic example let us prove that the sequence $\{x_i\}_{i=1}^{\infty}$ with $x_i = \frac{i-1}{i}$ is a Cauchy sequence. We have for $j > i$

$$\left|\frac{i-1}{i} - \frac{j-1}{j}\right| = \left|\frac{(i-1)j - i(j-1)}{ij}\right| = \left|\frac{i-j}{ij}\right| \leq \frac{1}{i}.$$

For a given $\epsilon > 0$, we now choose the natural number $N \geq 1/\epsilon$, so that $\frac{1}{N+1} \leq \epsilon$, in which case we have

$$\left|\frac{i-1}{i} - \frac{j-1}{j}\right| \leq \epsilon \quad \text{if } i, j \geq N.$$

This shows that $\{x_i\}$ with $x_i = \frac{i-1}{i}$ is a Cauchy sequence, and thus converges to a limit $\lim_{i\to\infty} x_i$. We proved above that $\lim_{i\to\infty} x_i = 1$.

14.6 Computing $\sqrt{2}$ by the Deca-section algorithm

We now describe a variation of the Bisection algorithm for $x^2 - 2 = 0$ called the Deca-section algorithm. Like the Bisection algorithm, the Deca-section algorith produces a sequence of numbers $\{x_i\}_{i=0}^{\infty}$ that converges to $\sqrt{2}$. In the Deca-section algorithm, the element x_i agrees with $\sqrt{2}$ to i decimal places, and thus the rate of convergence is easy to grip.

The Deca-section algorithm looks the same as the bisection algorithm except that at each step the current interval is divided into 10 subintervals instead of 2. We start again with $f(x) = x^2 - 2$ and $x_0 = 1$ and $X_0 = 2$ so that $f(x_0) < 0$ and $f(X_0) > 0$. Now we compute the value of f at the intermediate rational points 1.1, 1.2, $\cdots$, 1.9 and then choose two consecutive numbers x_1 and X_1 with $f(x_1) < 0$ and $f(X_1) > 0$. There has to be two such consecutive points because we know that $f(x_0) = f(1) < 0$ and then either $f(y) < 0$ for all $y = 1.1, 1.2, \cdots, 1.9$ at which point $f(2) > 0$, so we set $x_1 = 1.9$ and $X_1 = 2$, or $f(y) > 0$ at some intermediate point. We find that this gives $x_1 = 1.4$ and $X_1 = 1.5$. Now we continue the process by evaluating f at the rational numbers 1.41, 1.42, $\cdots$, 1.49, and then choosing two consecutive numbers x_2 and X_2 with $f(x_2) < 0$ and $f(X_2) > 0$. This gives $x_2 = 1.41$ and $X_2 = 1.42$. Then we work on the third, fourth, fifth, $\cdots$ decimal places in order, obtaining two sequences $\{x_i\}_{i=0}^{\infty}$ and $\{X_i\}_{i=0}^{\infty}$ both converging to $\sqrt{2}$. We show the first 14 steps computed using a *MATLAB*© m-file implementation of this algorithm in Fig. 14.3.

i	x_i	X_i
0	1.00000000000000	2.00000000000000
1	1.40000000000000	1.50000000000000
2	1.41000000000000	1.42000000000000
3	1.41400000000000	1.41500000000000
4	1.41420000000000	1.41430000000000
5	1.41421000000000	1.41422000000000
6	1.41421300000000	1.41421400000000
7	1.41421350000000	1.41421360000000
8	1.41421356000000	1.41421357000000
9	1.41421356200000	1.41421356300000
10	1.41421356230000	1.41421356240000
11	1.41421356237000	1.41421356238000
12	1.41421356237300	1.41421356237400
13	1.41421356237300	1.41421356237310
14	1.41421356237309	1.41421356237310

FIGURE 14.3. 14 steps of the deca-section algorithm.

By construction

$$|x_i - X_i| \leq 10^{-i},$$

and also

$$|x_i - x_j| \leq 10^{-i} \quad \text{for } j \geq i. \tag{14.7}$$

The inequality (14.7) implies that $\{x_i\}$ is a Cauchy sequence and thus determines an infinite decimal expansion. Since in the Deca-section algorithm, we gain one decimal per step, we may identify element x_i of the sequence with the truncated decimal expansion with i decimals. In this case there is thus a very simple connection between the Cauchy sequence and the decimal expansion.

Chapter 14 Problems

14.1. Use the *evalf* function in *MAPLE*© to compute $\sqrt{2}$ to 1000 places and then square the result and compare to 2.

14.2. (a) Show that $\sqrt{3}$ (see Problem 3.5) is irrational. Hint: use a powerful mathematical technique: try to copy a proof you already know. (b) Do the same for $\sqrt{a}$ where a is any prime number.

14.3. Specify three different irrational numbers using the digits 3 and 4.

14.4. Program the Bisection algorithm. Write down the output for 30 steps starting with: (a) $x_0 = 1$ and $X_0 = 2$, (b) $x_0 = 0$ and $X_0 = 2$, (c) $x_0 = 1$ and $X_0 = 3$, (d) $x_0 = 1$ and $X_0 = 20$. Compare the accuracy of the methods at each step by comparing the values of x_i versus the decimal expansion of $\sqrt{2}$ given above. Explain why there is a difference in accuracy resulting from the different initial values.

14.5. (a) Use the program in Problem 14.4 and write down the output for 40 steps using $x_0 = 1$ and $X_0 = 2$. (b) Describe anything you notice about the last 10 values x_i and X_i. (c) Explain what you see. (Hint: consider floating point representation on the computer you use.)

14.6. Using the results in Problem 14.4(a), make plots of: (a) $|X_i - x_i|$ versus i (b) $|x_i - x_{i-1}|$ versus i; and (c) $|f(x_i)|$ versus i. In each case, determine if the quantity decreases by a factor of 1/2 after each step.

14.7. For the sequence $\{X_i\}_{i=0}^{\infty}$ produced by the bisection algorithm, show that $\sqrt{2} = \lim_{i\to\infty} X_i$.

14.8. Solve the equations $x^2 = 3$ and $x^3 = 2$ using the bisection algorithm. Also, make the algorithm find the negative root of $x^2 = 3$.

14.9. Show that if $a < 0$ and $b > 0$ then $b - a < c$ implies $|b| < c$ and $|a| < c$.

14.10. (a) Write down an algorithm for Deca-section. (b) Program the algorithm in (a) and then compute 16 steps using $x_0 = 0$ and $X_0 = 2$.

14.11. Show that (16.5) is valid.

14.12. (a) Construct a "trisection" algorithm; (b) implement the trisection algorithm and compute 30 steps using $x_0 = 0$ and $X_0 = 2$; (c) show that the tridiagonal algorithm determines a decimal expansion and call this $\bar{x}$; (d) Show that $\bar{x} = \sqrt{2}$; (e) get an estimate on $|x_i - \bar{x}|$.

14.13. Compute the cost of the tridiagonal algorithm from Problem 14.12 and compare to the costs of the bistable and deca-section methods.

14.14. Use the bisection code from Problem 14.4 to compute $\sqrt{3}$ (recall Problem 3.5). Hint: $1 < \sqrt{3} < 2$.

This is page 207
Printer: Opaque this

15
Real numbers

I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind: it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be. (Kelvin 1889)

Vattnet drar sig tillbaka
stenarna blir synliga.

Det var länge sen sist.
De har egentligen inte förändrats.

De gamla stenarna.

(Brunnen, Lars Gustafsson, 1977)

15.1 Introduction

We are now ready to introduce the concept of a *real number*. We shall view a real number as being specified by an *infinite decimal expansion* of the form

$$\pm p_m \cdots p_0.q_1 q_2 q_3 \cdots$$

with a neverending list of decimals q_1, q_2, ..., where each one of the p_i and q_j are one of the 10 digits $0, 1, ..., 9$. We met the decimal expansion

1.4142135623.... of $\sqrt{2}$ above. The corresponding sequence $\{x_i\}_{i=1}^{\infty}$ of truncated decimal expansions is given by the rational numbers

$$x_i = \pm p_m \cdots p_0.q_1 \cdots q_i = \pm(p_m 10^m + \cdots + q_i 10^{-i}).$$

We have for $j > i$,

$$|x_i - x_j| = |0.0 \cdots 0 q_{i+1} \cdots q_j| \leq 10^{-i}. \tag{15.1}$$

We conclude that the sequence $\{x_i\}_{i=1}^{\infty}$ of truncated decimal expansions of the infinite decimal expansion $\pm p_m \cdots p_0.q_1 q_2 q_3 \cdots$, is a Cauchy sequence of rational numbers.

More generally, we know from the discussion in Chapter Sequences and Limits, that any Cauchy sequence of rational numbers specifies an infinite decimal expansion and thus a Cauchy sequence of rational numbers specifies a real number. We may thus view a real number as being specified by an infinite decimal expansion, or by a Cauchy sequence of rational numbers. Note that we use the semantic trick of referring to an infinite decimal expansion as one real number.

We divide real numbers into two types: *rational numbers* with periodic decimal expansions and *irrational numbers* with non-periodic decimal expansions. Note that we may naturally include rational numbers with finitely many nonzero decimals, like 0.25, as particular periodic infinite decimal expansions with all the decimals $q_i = 0$ for i sufficiently large.

We say that the infinite decimal expansion $\pm p_m \cdots p_0.q_1 q_2 q_3 \cdots$ specifies the *real number* $x = \pm p_m \cdots p_0.q_1 q_2 q_3 \cdots$, and we agree to write

$$\lim_{i \to \infty} x_i = x, \tag{15.2}$$

where $\{x_i\}_{i=1}^{\infty}$ is the corresponding sequence of truncated decimal expansions of x. If $x = \pm p_m \cdots p_0.q_1 q_2 q_3 \cdots$ is a periodic expansion, that is if x is a rational number, this agrees with our earlier definition from Chapter Sequences and Limits of the limit of the sequence $\{x_i\}_{i=1}^{\infty}$ of truncated decimal expansions of x. For example, we recall that

$$\frac{10}{9} = \lim_{i \to \infty} x_i, \quad \text{where } x_i = 1.11 \cdots 1_i.$$

If $x = \pm p_m \cdots p_0.q_1 q_2 q_3 \cdots$ is non-periodic, that is, if x is an *irrational number*, then (15.2) serves as a definition, where the real number is specified by the decimal expansion $x = \pm p_m \cdots p_0.q_1 q_2 q_3 \cdots$, that is the real number x specified by the Cauchy sequence $\{x_i\}_{i=1}^{\infty}$ of truncated decimal expansions of x, is *denoted* by $\lim_{i \to \infty} x_i$. Alternatively, (15.2) serves to denote the limit $\lim_{i \to \infty} x_i$ by $\pm p_m \cdots p_0.q_1 q_2 q_3 \cdots$.

For the sequence $\{x_i\}_{i=1}^{\infty}$ generated by the Bisection algorithm applied to the equation $x^2 - 2 = 0$, we decided to write $\sqrt{2} = \lim_{i \to \infty} x_i$, and thus we may write $\sqrt{2} = 1.412...$, with 1.412..., denoting the infinite decimal expansion given by the Bisection algorithm.

We shall now specify how to compute with real numbers defined in this way. In particular, we shall specify how to add, subtract, multiply and divide real numbers. Of course we will do this so that it extends our experience in computing with rational numbers. This will complete our process of extending the natural numbers to obtain first the integers and then the rational numbers, and finally the real numbers.

We denote by $\mathbb{R}$ the set of all possible real numbers, that is the set of all possible infinite decimal expansions. We discuss this definition in the Chapter Quarrell below.

15.2 Adding and subtracting real numbers

To exhibit the main concern, consider the problem of adding two real numbers x and $\bar{x}$ specified by the decimal expansions

$$x = \pm p_m \cdots p_0.q_1q_2q_3 \cdots = \lim_{i\to\infty} x_i,$$
$$\bar{x} = \pm \bar{p}_m \cdots \bar{p}_0.\bar{q}_1\bar{q}_2\bar{q}_3 \cdots = \lim_{i\to\infty} \bar{x}_i,$$

with corresponding truncated decimal expansions

$$x_i = \pm p_m \cdots p_0.q_1 \cdots q_i,$$
$$\bar{x}_i = \pm \bar{p}_m \cdots \bar{p}_0.\bar{q}_1 \cdots \bar{q}_i.$$

We know how to add x_i and $\bar{x}_i$: we then start from the right and add the decimals q_i and $\bar{q}_i$, and get a new ith decimal and possibly a carry-over digit to be added to the sum of the next digits q_{i-1} and $\bar{q}_{i-1}$, and so on. The important thing to notice is that we start from the right (smallest decimal) and move to the left (larger decimals).

Now, trying to add the two infinite sequences $x = \pm p_m \cdots p_0.q_1q_2q_3 \cdots$ and $\bar{x} = \pm \bar{p}_m \cdots \bar{p}_0.\bar{q}_1\bar{q}_2\bar{q}_3 \cdots$ in the same way by starting from the right, we run into a difficulty because there is no far right decimal to start with. So, what can we do?

Well, the natural way out is of course to consider the sequence $\{y_i\}$ generated by $y_i = x_i + \bar{x}_i$. Since both $\{x_i\}$ and $\{\bar{x}_i\}$ are Cauchy sequences, it follows that $\{y_i\}$ is also a Cauchy sequence, and thus defines a decimal expansion and thus defines a real number. Of course, the right thing is then to define

$$x + \bar{x} = \lim_{i\to\infty} y_i = \lim_{i\to\infty} (x_i + \bar{x}_i).$$

This corresponds to the formula

$$\lim_{i\to\infty} x_i + \lim_{i\to\infty} \bar{x}_i = \lim_{i\to\infty} (x_i + \bar{x}_i).$$

We give a concrete example: To compute the sum of

$$x = \sqrt{2} = 1.4142135623730950488\cdots$$

and

$$\bar{x} = \frac{1043}{439} = 2.3758542141230068337\cdots,$$

we compute $y_i = x_i + \bar{x}_i$ for $i = 1, 2, ...$, which defines the decimal expansion of $x + \bar{x}$, see Fig. 15.1. We may notice that occasionally adding two digits

i	$x_i + \bar{x}_i$
1	3
2	3.7
3	3.78
4	3.789
5	3.7900*
6	3.79006
7	3.790067
8	3.7900677
9	3.79006777
10	3.790067776
11	3.7900677764
12	3.79006777649
13	3.790067776496
14	3.7900677764960
15	3.79006777649609
16	3.790067776496101*
17	3.7900677764961018
18	3.79006777649610187
19	3.790067776496101881*
20	3.7900677764961018825*
$\vdots$	$\vdots$

FIGURE 15.1. Computing the decimal expansion of $\sqrt{2} + 1043/439$ by using the truncated decimal sequences. Note the changes in the digits marked by the $*$ where adding the new digits affects previous digits.

affects the digits to the left, as in $0.9999 + 0.0001 = 1.000$.

Similarly, the difference $x - \bar{x}$ of two real numbers $x = \lim_{i\to\infty} x_i$ and $\bar{x} = \lim_{i\to\infty} \bar{x}_i$ is of course defined by

$$x - \bar{x} = \lim_{i\to\infty} (x_i - \bar{x}_i).$$

15.3 Generalization to $f(x, \bar{x})$ with f Lipschitz

We now generalize to other combinations of real numbers than addition. Suppose we want to combine x and $\bar{x}$ to a certain quantity $f(x, \bar{x})$ depending on x and $\bar{x}$, where x and $\bar{x}$ are real numbers. For example, we may choose $f(x, \bar{x}) = x + \bar{x}$, corresponding to determining the sum $x + \bar{x}$ of two real numbers x and $\bar{x}$ or $f(x, \bar{x}) = x\bar{x}$ corresponding to multiplying x and $\bar{x}$.

To be able to define $f(x, \bar{x})$ following the idea used in the case $f(x, \bar{x}) = x + \bar{x}$, we suppose that $f : \mathbb{Q} \times \mathbb{Q} \to \mathbb{Q}$ is Lipschitz continuous. This is a very crucial assumption and our focus on the concept of Lipschitz continuity is largely motivated by its use in the present context.

We know from Chapter Sequences and Limits that if $x = \lim_{i\to\infty} x_i$ and $\bar{x} = \lim_{i\to\infty} \bar{x}_i$ are rational, then

$$f(x, \bar{x}) = f(\lim_{i\to\infty} x_i, \lim_{i\to\infty} \bar{x}_i) = \lim_{i\to\infty} f(x_i, \bar{x}_i)$$

If $x = \lim_{i\to\infty} x_i$ and $\bar{x} = \lim_{i\to\infty} \bar{x}_i$ are irrational, we simply decide to use this formula to *define* the real number $f(x, \bar{x})$. This is possible, because $\{f(x_i, \bar{x}_i)\}$ is a Cauchy sequence and thus defines a real number. Note that $\{f(x_i, \bar{x}_i)\}$ is a Cauchy sequence because $\{x_i\}$ and $\{\bar{x}_i\}$ are both Cauchy sequences and $f : \mathbb{Q} \times \mathbb{Q} \to \mathbb{Q}$ is Lipschitz continuous. The formula containing this crucial information is

$$|f(x_i, \bar{x}_i) - f(x_j, \bar{x}_j)| \leq L(|x_i - x_j| + |\bar{x}_i - \bar{x}_j|)$$

where L is the Lipschitz constant of f.

Applying this reasoning to the case $f : \mathbb{Q} \times \mathbb{Q} \to \mathbb{Q}$ with $f(x, \bar{x}) = x + \bar{x}$, which is Lipschitz continuous with Lipschitz constant $L = 1$, we define the sum $x + \bar{x}$ of two real numbers $x = \lim_{i\to\infty} x_i$ and $\bar{x} = \lim_{i\to\infty} \bar{x}_i$ by

$$x + \bar{x} = \lim_{i\to\infty} (x_i + \bar{x}_i),$$

that is

$$\lim_{i\to\infty} x_i + \lim_{i\to\infty} \bar{x}_i = \lim_{i\to\infty} (x_i + \bar{x}_i). \tag{15.3}$$

This is exactly what we did above.

We repeat, the important formula is

$$f(\lim_{i\to\infty} x_i, \lim_{i\to\infty} \bar{x}_i) = \lim_{i\to\infty} f(x_i, \bar{x}_i),$$

which we already know for $x = \lim_{i\to\infty} x_i$ and $\bar{x} = \lim_{i\to\infty} \bar{x}_i$ rational and which defines $f(x, \bar{x})$ for x or $\bar{x}$ irrational. We also repeat that the Lipschitz continuity of f is crucial.

We may directly extend to Lipschitz functions $f : I \times J \to \mathbb{Q}$, where I and J are intervals of $\mathbb{Q}$, under the assumption that $x_i \in I$ and $\bar{x}_i \in J$ for $i = 1, 2, ...$

15.4 Multiplying and dividing real numbers

The function $f(x, \bar{x}) = x\bar{x}$ is Lipschitz continuous for $x \in I$ and $\bar{x} \in J$, where I and J are bounded intervals of $\mathbb{Q}$. We may thus define the product $x\bar{x}$ of two real numbers $x = \lim_{i\to\infty} x_i$ and $\bar{x} = \lim_{i\to\infty} \bar{x}_i$ as follows:

$$x\bar{x} = \lim_{i\to\infty} x_i\bar{x}_i.$$

The function $f(x, \bar{x}) = \frac{x}{\bar{x}}$ is Lipschitz continuous for $x \in I$ and $\bar{x} \in J$, if I and J are bounded intervals of $\mathbb{Q}$ and J is bounded away from 0. We may thus define the quotient $\frac{x}{\bar{x}}$ of two real numbers $x = \lim_{i\to\infty} x_i$ and $\bar{x} = \lim_{i\to\infty} \bar{x}_i$ with $\bar{x} \neq 0$ by

$$\frac{x}{\bar{x}} = \lim_{i\to\infty} \frac{x_i}{\bar{x}_i}.$$

15.5 The absolute value

The function $f(x) = |x|$ is Lipschitz continuous on $\mathbb{Q}$. We may thus define the absolute value $|x|$ of a real number $x = \lim_{i\to\infty} x_i$ by

$$|x| = \lim_{i\to\infty} |x_i|.$$

If $\{x_i\}$ is the sequence of truncated decimal expansions of $x = \lim_{i\to\infty} x_i$, then by (15.1) we have $|x_j - x_i| \leq 10^{-i}$ for $j > i$, and thus taking the limit as j tends to infinity,

$$|x - x_i| \leq 10^{-i} \quad \text{for } i = 1, 2, ... \tag{15.4}$$

15.6 Comparing two real numbers

Let $x = \lim_{i\to\infty} x_i$ and $\bar{x} = \lim_{i\to\infty} \bar{x}_i$ be two real numbers with corresponding sequences of truncated decimal expansions $\{x_i\}_{i=1}^{\infty}$ and $\{\bar{x}_i\}_{i=1}^{\infty}$. How can we tell if $x = \bar{x}$? Is it necessary that $x_i = \bar{x}_i$ for all i? Not quite. For example, consider the two numbers $x = 0.99999\cdots$ and $\bar{x} = 1.0000\cdots$. In fact, it is natural to give a little more freedom and say that $x = \bar{x}$ if and only if

$$|x_i - \bar{x}_i| \leq 10^{-i} \quad \text{for } i = 1, 2, ... \tag{15.5}$$

This condition is clearly sufficient to motivate to write $x = \bar{x}$, since the difference $|x_i - \bar{x}_i|$ becomes as small as we please by taking i large enough. In other words, we have

$$|x - \bar{x}| = \lim_{i \to \infty} |x_i - \bar{x}_i| = 0,$$

so $x = \bar{x}$.

Conversely if (15.5) does not hold, then there is a positive ϵ and i such that

$$x_i - \bar{x}_i > 10^{-i} + \epsilon \quad \text{or} \quad x_i - \bar{x}_i < 10^{-i} - \epsilon.$$

Since $|x_i - x_j| \leq 10^{-i}$ for $j > i$, we must then have

$$x_j - \bar{x}_j > \epsilon \quad \text{or} \quad x_j - \bar{x}_j < -\epsilon \quad \text{for } j > i$$

and thus taking the limit as j tends to infinity

$$x - \bar{x} \geq \epsilon \quad \text{or} \quad x - \bar{x} \leq -\epsilon.$$

We conclude that two real numbers x and $\bar{x}$ either satisfy $x = \bar{x}$, or $x > \bar{x}$ or $x < \bar{x}$.

This conclusion, however, hides a subtle point. To know if two real numbers are equal or not, may require a complete knowledge of the decimal expansions, which may not be realistic. For example, suppose we set $x = 10^{-p}$, where p is the decimal position of the start of the first sequence of 59 decimals all equal to 1 in the decimal expansion of $\sqrt{2}$. To complete the definition of x, we set $x = 0$ if there is no such p. How are we to know if $x > 0$ or $x = 0$, unless we happen to find that sequence of 59 decimals all equal to 1 among say the first 10^{50} decimals, or whatever number of decimals of $\sqrt{2}$ we can think of possibly computing. In a case like this, it seems more reasonable to say that we cannot know if $x = 0$ or $x > 0$.

15.7 Summary of arithmetics with real numbers

With these definitions, we can easily show that the usual commutative, distributive, and associative rules for rational numbers all hold for real numbers. For example, addition is commutative since

$$x + \bar{x} = \lim_{i \to \infty} (x_i + \bar{x}_i) = \lim_{i \to \infty} (\bar{x}_i + x_i) = \bar{x} + x.$$

15.8 Why $\sqrt{2}\sqrt{2}$ equals 2

Let $\{x_i\}$ and $\{X_i\}$ be the sequences given by the Bisection Algorithm applied to the equation $x^2 = 2$ constructed above. We have defined

$$\sqrt{2} = \lim_{i \to \infty} x_i, \tag{15.6}$$

that is, we denote by $\sqrt{2}$ the infinite non-periodic decimal expansion given by the Bisection algorithm applied to the equation $x^2 = 2$.

We now verify that $\sqrt{2}\sqrt{2} = 2$, which we left open above. By the definition of multiplication of real numbers, we have

$$\sqrt{2}\sqrt{2} = \lim_{i\to\infty} x_i^2, \tag{15.7}$$

and we thus need to show that

$$\lim_{i\to\infty} x_i^2 = 2 \tag{15.8}$$

To prove this fact, we use the Lipschitz continuity of the function $x \to x^2$ on $[0,2]$ with Lipschitz constant $L = 4$, to see that

$$|(x_i)^2 - (X_i)^2| \leq 4|x_i - X_i| \leq 2^{-i+2}.$$

where we use the inequality $|x_i - X_i| \leq 2^{-i}$. By construction $x_i^2 < 2 < X_i^2$, and thus in fact

$$|x_i^2 - 2| \leq 2^{-i+2}$$

which shows that

$$\lim_{i\to\infty} (x_i)^2 = 2$$

and (15.8) follows.

We summarize the approach used to compute and define $\sqrt{2}$ as follows:

- We use the Bisection Algorithm applied to the equation $x^2 = 2$ to define a sequence of rational numbers $\{x_i\}_{i=0}^{\infty}$ that converges to a limit, which we denote by $\sqrt{2} = \lim_{i\to\infty} x_i$.
- We define $\sqrt{2}\sqrt{2} = \lim_{i\to\infty}(x_i)^2$.
- We show that $\lim_{i\to\infty}(x_i)^2 = 2$.
- We conclude that $\sqrt{2}\sqrt{2} = 2$ which means that $\sqrt{2}$ solves the equation $x^2 = 2$.

15.9 A reflection on the nature of $\sqrt{2}$

We may now return to comparing the following two definitions of $\sqrt{2}$:

1. $\sqrt{2}$ is "that thing" which when squared is equal to 2
2. $\sqrt{2}$ is the name of the decimal expansion given by the sequence $\{x_i\}_{i=1}^{\infty}$ generated by the Bisection algorithm for the equation $x^2 = 2$, which with a suitable definition of multiplication satisfies $\sqrt{2}\sqrt{2} = 2$.

This is analogous to the following two definitions of $\frac{1}{2}$:

1. $\frac{1}{2}$ is "that thing" which when multiplied by 2 equals 1
2. $\frac{1}{2}$ is the ordered pair $(1, 2)$ which with a suitable definition of multiplication satisfies the equation $(2, 1) \times (1, 2) = (1, 1)$.

We conclude that in both cases the meaning 1. could be criticized for being unclear in the sense that no clue is given to what "that thing" is in terms of already known things, and that the definition appears circular and eventually seems to be just a play with words. We conclude that only the definition 2. has a solid constructive basis, although we may intuitively use 1. when we *think*.

Occasionally, we can do computations including $\sqrt{2}$, where we only need to use that $(\sqrt{2})^2 = 2$, and we do not need the decimal expansion of $\sqrt{2}$. For example, we can verify that $(\sqrt{2})^4 = 4$ by only using that $(\sqrt{2})^2 = 2$ without knowing a single decimal of $\sqrt{2}$. In this case we just use $\sqrt{2}$ as a *symbol* for "that thing which squared equals 2". It is rare that this kind of symbolic manipulation only, leads to the end and gives a definite answer.

We note that the fact that $\sqrt{2}$ solves the equation $x^2 = 2$ includes some kind of convention or agreement or definition. What we actually did was to show that the truncated decimal expansions of $\sqrt{2}$ when squared could be made arbitrarily close to 2. We took this as a definition, or agreement, that $(\sqrt{2})^2 = 2$. Doing this, solved the dilemma of the Pythagoreans, and thus we may (almost) say that we solved the problem by *agreeing* that the problem did not exist. This may be the only way out in some (difficult) cases.

In fact, the standpoint of the famous philosopher Wittgenstein was that the only way to solve *philosophical problems* was to show (after much work) that in fact the problem at hand does not exist. The net result of this kind of reasoning would appear to be zero: first posing a problem and then showing that the problem does not exist. However, the process itself of coming to this conclusion would be considered as important by giving added insight, not so much the result. This approach also could be fruitful outside philosophy or mathematics.

15.10 Cauchy sequences of real numbers.

We may extend the notion of sequence and Cauchy sequence to real numbers. We say that $\{x_i\}_{i=1}^{\infty}$ is a sequence of real numbers if the elements x_i are real numbers. The definition of convergence is the same as for sequences of rational numbers. A sequence $\{x_i\}_{i=1}^{\infty}$ of real numbers converges to a real number x if for any $\epsilon > 0$ there is a natural number N such that $|x_i - x| < \epsilon$ for $i \geq N$ and we write $x = \lim_{i \to \infty} x_i$.

We say that a sequence $\{x_i\}_{i=1}^{\infty}$ of real numbers is a Cauchy sequence if for all $\epsilon > 0$ there is a natural number N such that

$$|x_i - x_j| \leq \epsilon \quad \text{for} \quad i, j \geq N. \tag{15.9}$$

If $\{x_i\}_{i=1}^{\infty}$ is a converging sequence of real numbers with limit $x = \lim_{i\to\infty} x_i$, then by the Triangle Inequality,

$$|x_i - x_j| \leq |x - x_i| + |x - x_j|,$$

where we wrote $x_i - x_j = x_i - x + x - x_j$. This proves that $\{x_i\}_{i=1}^{\infty}$ is a Cauchy sequence. We state this (obvious) result as a theorem.

Theorem 15.1 *A converging sequence of real numbers is a Cauchy sequence of real numbers.*

A Cauchy sequence of real numbers determines a decimal expansion just in the same way as a sequence of rational numbers does. We may assume, possibly by deleting elements and changing the indexing, that a Cauchy sequence of real numbers satisfies $|x_i - x_j| \leq 2^{-i} \quad \text{for} \quad j \geq i$.

We conclude that a Cauchy sequence of real numbers converges to a real number. This is a fundamental result about real numbers which we state as a theorem.

Theorem 15.2 *A Cauchy sequence of real numbers converges to a unique real number.*

The use of Cauchy sequences has been popular in mathematics since the days of the great mathematician Cauchy in the first half of the 19th century. Cauchy was a teacher at Ecole Polytechnique in Paris, which was created by Napoleon and became a model for technical universities all over Europe (Chalmers 1829, Helsinki 1849,....). Cauchy's reform of the engineering Calculus course including his famous Cours d'Analyse also became a model, which permeates much of the Calculus teaching still today.

15.11 Extension from $f : \mathbb{Q} \to \mathbb{Q}$ to $f : \mathbb{R} \to \mathbb{R}$

In this section, we show how to *extend* a given Lipschitz continuous function $f : \mathbb{Q} \to \mathbb{Q}$, to a function $f : \mathbb{R} \to \mathbb{R}$. We thus assume that $f(x)$ is defined for x rational, and that $f(x)$ is a rational number, and we shall now show how to define $f(x)$ for x irrational. We shall see that the Lipschitz continuity is crucial in this extension process. In fact, much of the motivation for introducing the concept of Lipschitz continuity, comes from its use in this context.

We have already met the basic issues when defining how to compute with real numbers, and we follow the same idea for a general function $f : \mathbb{Q} \to \mathbb{Q}$.

If $x = \lim_{i\to\infty} x_i$ is an irrational real number, with the sequence $\{x_i\}_{i=1}^\infty$ being the truncated decimal expansions of x, we define $f(x)$ to be the real number defined by

$$f(x) = \lim_{i\to\infty} f(x_i). \tag{15.10}$$

Note that by the Lipschitz continuity of $f(x)$ with Lipschitz constant L, we have

$$|f(x_i) - f(x_j)| \le L|x_i - x_j|,$$

which shows that the sequence $\{f(x_i)\}_{i=1}^\infty$ is a Cauchy sequence, because $\{x_i\}_{i=1}^\infty$ is a Cauchy sequence Thus $\lim_{i\to\infty} f(x_i)$ exists and defines a real number $f(x)$. This defines $f : \mathbb{R} \to \mathbb{R}$, and we say that this function is the *extension* of $f : \mathbb{Q} \to \mathbb{Q}$, from the rational numbers $\mathbb{Q}$ to the real numbers $\mathbb{R}$.

Sinmilarly, we can generalize and extend a Lipschitz continuous function $f : I \to \mathbb{Q}$, where $I = \{x \in \mathbb{Q} : a \le x \le b\}$ is an interval of rational numbers, to a function $f : J \to \mathbb{R}$, where $J = \{x \in \mathbb{R} : a \le x \le b\}$ is the corresponding interval of real numbers. Evidently, the extended function $f : J \to \mathbb{R}$ satisfies:

$$f(\lim_{i\to\infty} x_i) = \lim_{i\to\infty} f(x_i), \tag{15.11}$$

for any convergent sequence $\{x_i\}$ in J (with automatically $\lim_{i\to\infty} x_i \in J$ because J is closed).

15.12 Lipschitz continuity of extended functions

If $f : \mathbb{Q} \to \mathbb{Q}$ is Lipschitz continuous with Lipschitz constant L_f, then its extension $f : \mathbb{R} \to \mathbb{R}$ is also Lipschitz continuous with the same Lipschitz constant L_f. This is because if $x = \lim_{i\to\infty} x_i$ and $y = \lim_{i\to\infty} y_i$, then

$$|f(x) - f(y)| = \left|\lim_{i\to\infty} (f(x_i) - f(y_i))\right| \le L \lim_{i\to\infty} |x_i - y_i| = L|x - y|.$$

It is now straightforward to show that the properties of Lipschitz continuous functions $f : \mathbb{Q} \to \mathbb{Q}$ stated above hold for the corresponding extended functions $f : \mathbb{R} \to \mathbb{R}$. We summarize in the following theorem

Theorem 15.3 *A Lipschitz continuous function $f : I \to \mathbb{R}$, where $I = [a, b]$ is an interval of real numbers, satisfies:*

$$f(\lim_{i\to\infty} x_i) = \lim_{i\to\infty} f(x_i), \tag{15.12}$$

for any convergent sequence $\{x_i\}$ in I. If $f : I \to \mathbb{R}$ and $g : I \to \mathbb{R}$ are Lipschitz continuous, and α and β are real numbers, then the linear combination $\alpha f(x) + \beta g(x)$ is Lipschitz continuous on I. If the interval

I is bounded, then $f(x)$ and $g(x)$ are bounded and $f(x)g(x)$ is Lipschitz continuous on I. If I is bounded and moreover $|g(x)| \geq c > 0$ for all x in I, where c is some constant, then $f(x)/g(x)$ is Lipschitz continuous on I.

EXAMPLE 15.1. We can extend any polynomial to be defined on the real numbers. This is possible because a polynomial is Lipschitz continuous on any bounded interval of rational numbers.

EXAMPLE 15.2. The previous example means that we can extend $f(x) = x^n$ to the real numbers for any integer n. We can also show that $f(x) = x^{-n}$ is Lipschitz continuous on any closed interval of rational numbers that does not contain 0. Therefore $f(x) = x^n$ can be extended to the real numbers, where n is any integer provided that when $n < 0$, $x \neq 0$.

15.13 Graphing functions $f : \mathbb{R} \to \mathbb{R}$

Graphing a function $f : \mathbb{R} \to \mathbb{R}$ follows the same principles as graphing a function $f : \mathbb{Q} \to \mathbb{Q}$.

15.14 Extending a Lipschitz continuous function

Suppose $f : (a, b] \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant L_f on the half-open interval $(a, b]$, but that the value of $f(a)$ has not been defined. Is there a way to define $f(a)$ so that the extended function $f : [a, b] \to \mathbb{R}$ is Lipschitz continuous? Yes, there is. To see this we let $\{x_i\}_{i=1}^{\infty}$ be a sequence of real numbers in $(a, b]$ converging to a, that is $\lim_{i\to\infty} x_i = a$. The sequence $\{x_i\}_{i=1}^{\infty}$ is a Cauchy sequence, and because $f(x)$ is Lipschitz continuous on $(a, b]$, so is the sequence $\{f(x_i)\}_{i=1}^{\infty}$, and thus $\lim_{i\to\infty} f(x_i)$ exists and we may then define $f(a) = \lim_{i\to\infty} f(x_i)$. It follows readily that the extended function $f : [a, b] \to \mathbb{R}$ is Lipschitz continuous with the same Lipschitz constant.

We give an application of this idea arising when considering quotients of two functions. Clearly, we must avoid points at which the denominator is zero and the numerator is nonzero. However, if both the numerator and denominator are zero at a point, the function can be extended to that point if the quotient function is Lipschitz continuous off the point. We give first a “trivial” example.

EXAMPLE 15.3. Consider the quotient

$$\frac{x-1}{x-1}$$

with domain $\{x \in \mathbb{R} : x \neq 1\}$. Since

$$x - 1 = 1 \times (x - 1) \tag{15.13}$$

for all x, it is natural to "divide" the polynomials to get

$$\frac{x-1}{x-1} = 1. \tag{15.14}$$

However, the domain of the constant function 1 is $\mathbb{R}$ so the left- and right-hand sides of (15.14) have different domains and therefore must represent different functions. We plot the two functions in Fig. 15.2. We

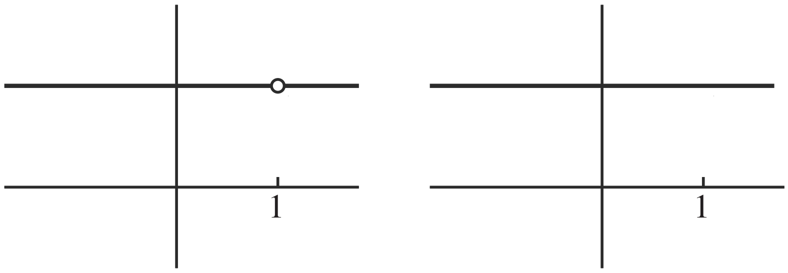


FIGURE 15.2. Plots of $(x-1)/(x-1)$ on the left and 1 on the right.

see that the two functions agree at every point except for the "missing" point $x = 1$.

EXAMPLE 15.4. Since $x^2 - 2x - 3 = (x-3)(x+1)$, we have for $x \neq 3$ that

$$\frac{x^2 - 2x - 3}{x - 3} = x + 1.$$

The function $(x^2 - 2x - 3)/(x - 3)$ defined for $\{x \in \mathbb{R} : x \neq 3\}$ may be extended to the function $x + 1$ defined for all x in $\mathbb{R}$.

Note that the fact that two functions f_1 and f_2 are zero at the same points does not mean that we can automatically replace their quotient by a function defined at all points.

EXAMPLE 15.5. The function

$$\frac{x-1}{(x-1)^2},$$

defined for $\{x \in \mathbb{R} : x \neq 1\}$, is equal to the function $1/(x-1)$ also defined on $\{x \in \mathbb{R} : x \neq 1\}$, which cannot be extended to $x = 1$.

15.15 Intervals of real numbers

Let a and b be two real numbers with $a < b$. The set of real numbers x such that $x > a$ and $x < b$, that is $\{x \in \mathbb{R} : a < x < b\}$, is called the *open interval* between a and b and is denoted by (a, b). Graphically we draw a thick line on the number line connecting little circles drawn at positions a and b. We illustrate in Fig. 15.3. The word "open" refers to

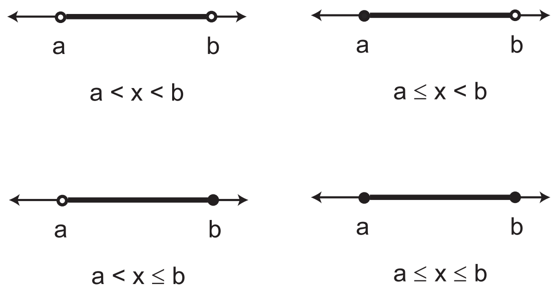


FIGURE 15.3. Intervals corresponding to the real numbers between two real numbers a and b. Note the use of a solid and closed circles in the four cases.

the strict inequality defining (a, b) and we use the curved parentheses "(" and the open circle on the number line to mark this. a and b are called the *endpoints* of the interval. An open interval does not contain its endpoints. The *closed interval* $[a, b]$ is the set $\{x : a \leq x \leq b\}$ and is denoted on the number line using solid circles. Note the use of square parentheses "[" when the inequalities are not strict. A closed interval does contain its endpoints. Finally, we can have *half-open intervals* with one end open and the other closed, such as $(a, b] = \{x : a < x \leq b\}$. See Fig. 15.3.

We also have "infinite" intervals such as $(-\infty, a) = \{x : x < a\}$ and $[b, \infty) = \{x : b \leq x\}$. We illustrate these in Fig. 15.4. With this notation,

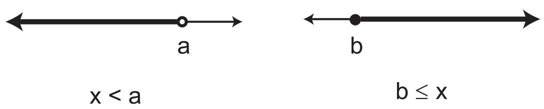


FIGURE 15.4. Infinite intervals $(-\infty, a)$ and $[b, \infty)$.

we denote the set of real numbers by $\mathbb{R} = (-\infty, \infty)$.

Clearly, we can now consider Lipschitz continuous functions $f : I \to \mathbb{R}$ defined on intervals I of $\mathbb{R}$.

15.16 What is $f(x)$ if x is irrational?

Note that if x is irrational, then the process of determining the sequence of truncated decimal expansions of x and $f(x)$ is carried out in parallel. The more decimals we have of x, the more decimals we get of $f(x)$. This is simply because $f(x) = \lim_{i\to\infty} f(x_i)$ with $\{x_i\}_{i=1}^{\infty}$ the sequence of truncated decimal expansions of x. This is obvious from Fig. 15.5. This means that the conventional idea of viewing $f(x)$ as a function of x comes into a new light. In the traditional way of thinking of a function $f(x)$, we think of x as given and then associating the value $f(x)$ to x. We may even write this as $x \to f(x)$ indicating that we go *from* x *to* $f(x)$.

However, we just noticed that when x is irrational, we cannot start from knowing all the decimals of x, and then determine $f(x)$. Instead, we determine successively the decimal expansions x_i and the corresponding function values $f(x_i)$, that is, we may write $x_i \to f(x_i)$ for $i = 1, 2, ...$, but not really $x \to f(x)$. We rather jump back and forth between approximations x_i of x and approximations $f(x_i)$ of $f(x)$. This means that we do not have exact knowledge of x when we compute $f(x)$. In order to make this process to be meaningful, we need the function $f(x)$ to be Lipschitz continuous. In this case, small changes in x cause small changes in $f(x)$, and the extension process is possible.

EXAMPLE 15.6. We evaluate $f(x) = .4x^3 - x$ for $x = \sqrt{2}$ using the truncated decimal sequence $\{x_i\}$ in Fig. 15.5.

i	x_i	$.4x_i^3 - x_i$
1	1	$-.6$
2	1.4	.0976
3	1.41	.1212884
4	1.414	.1308583776
5	1.4142	.1313383005152
6	1.41421	.1313623002245844
7	1.414213	.1313695002035846388
8	1.4142135	.13137070020305452415
9	1.41421356	.1313708442030479314744064
10	1.414213562	.1313708490030479221535281312
$\vdots$	$\vdots$	$\vdots$

FIGURE 15.5. Computing the decimal expansion of $f(\sqrt{2})$ for $f(x) = .4x^3 - x$ by using the truncated decimal sequence.

This leads to the idea that we can only talk about Lipschitz continuous functions. If some association of x-values to values $f(x)$ is not Lipschitz continuous, this association should not deserve to be called a function. We are thus led to the conclusion that *all functions are Lipschitz continuous* (more or less).

This statement would be shocking to many mathematicians, who are used to work with discontinuous functions day and night. In fact, in large parts of mathematics (e.g. integration theory), a lot of attention is payed to extremely discontinuous "functions", like the following popular one

$$\begin{aligned} f(x) &= 0 \quad \text{if } x \quad \text{is rational,} \\ f(x) &= 1 \quad \text{if } x \quad \text{is irrational.} \end{aligned}$$

Whatever this is, it is not a Lipschitz function, and thus from our perspective, we would not consider it to be a function at all. This is because for some arguments x it may be extremely difficult to know if x is rational or irrational, and then we would not know which of the vastly different function values $f(x) = 0$ or $f(x) = 1$ to choose. To be able to determine if x is rational or not, we may have to know the infinite decimal expansion of x, which may be impossible to us as human beings. For example, if we didn't know the smart argument showing that $x = \sqrt{2}$ cant be rational, we would not be able to tell from any truncated decimal expansion of $\sqrt{2}$ whether $f(x) = 0$ or $f(x) = 1$.

We would even get into trouble trying to define the following "function" $f(x)$

$$\begin{aligned} f(x) &= a \quad \text{if } x < \bar{x}, \\ f(x) &= b \quad \text{if } x \geq \bar{x}, \end{aligned}$$

with a "jump" at $\bar{x}$ from a value a to a different value b. If $\bar{x}$ is irrational, we may lack complete knowledge of all the decimals of $\bar{x}$, and it may be virtually impossible to determine for a given x if $x < \bar{x}$ or $x \geq \bar{x}$. It would be more natural to view the "function with a jump" as *two functions* composed of one Lipschitz function

$$f(x) = a \quad \text{if } x \leq \bar{x},$$

and another Lipschitz function

$$f(x) = b \quad \text{if } x \geq \bar{x},$$

with two possible values $a \neq b$ for $x = \bar{x}$: the value a from the left ($x \leq \bar{x}$), and the value b from the right ($x \geq \bar{x}$), see Fig. 15.6.

It thus seems that we have to reject the very idea that a function $f(x)$ can be discontinuous. This is because we cannot assume that we know x exactly, and thus we can only handle a situation where small changes in x causes small changes in $f(x)$, which is the essence of Lipschitz continuity. Instead we are led to handle functions with jumps as combinations of Lipschitz continuous functions with two possible values at the jumps, one value from the right and another value from the left.

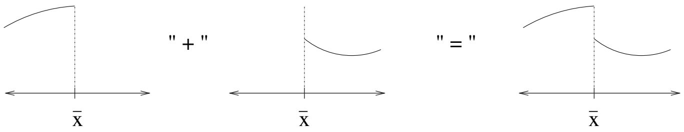


FIGURE 15.6. A "jump function" viewed as two functions

15.17 The usual definition of continuity vs Lipschitz continuity

As indicated, we use a definition of continuity (Lipschitz continuity), which differs from the usual definition met in most Calculus texts. We recall the basic property of a Lipschitz continuous function $f : I \to \mathbb{R}$:

$$f(\lim_{i\to\infty} x_i) = \lim_{i\to\infty} f(x_i), \tag{15.15}$$

for any convergent sequence $\{x_i\}$ in I with $\lim_{i\to\infty} x_i \in I$. Now, the standard definition of continuity of a function $f : I \to \mathbb{R}$ starts at the relation (15.15), and reads as follows: The function $f : I \to \mathbb{R}$ is said to be *continuous* on I (according to the standard definition) if (15.15) holds for any convergent sequence $\{x_i\}$ in I with $\lim_{i\to\infty} x_i \in I$. Apparently, a Lipschitz continuous function is continuous according to the standard definition, while the opposite implication does not have to be true. In other words, we use a somewhat more stringent definition than the standard one.

The standard definition satisfies a condition of maximality (attractive to many pure mathematicians), but suffers from an (often confusing) use of limits. In fact, the intuitive idea of "continuous dependence" of function values $f(x)$ of a real variable x, can be expressed as "$f(x)$ is close to $f(y)$ whenever x is close to y", of which Lipschitz continuity gives a quantitative precise formulation, while the connection in the standard definition is more farfetched. Right?

Chapter 15 Problems

15.1. Define a "sentence" to be any combination of 500 characters consisting of 26 letters and spaces lined up in a row. Compute (approximately) the number of possible sentences.

15.2. Suppose that x and y are two real numbers and $\{x_i\}$ and $\{y_i\}$ are the sequences generated by truncating their decimal expansions. Using (7.14) and (15.4), obtain estimates on (a) $|(x+y)-(x_i+y_i)|$ and (b) $|xy - x_iy_i|$. Hint: for (b), use that $xy - x_iy_i = (x-x_i)y + x_i(y-y_i)$, and explain why (15.4) implies that for i sufficiently big, $|x_i| \le |x| + 1$.

15.3. Find i as small as possible such that $|xy - x_i y_i| \le 10^{-6}$ if $x \approx 100$ and $y \approx 1$. Find i and j as small as possible such that $|xy - x_i y_j| \le 10^{-6}$

15.4. Let $x = .37373737\cdots$ and $y = \sqrt{2}$ and $\{x_i\}$ and $\{y_i\}$ be the sequences generated by truncating their decimal expansions. Compute the first 10 terms of the sequences defining $x + y$ and $y - x$ and the first 5 terms of the sequences defining xy and x/y. Hint: follow the example in Fig. 15.1.

15.5. Let x be the limit of the sequence $\left\{\frac{i}{i+1}\right\}$. Is $x < 1$?. Give a reason for your answer.

15.6. Let x be the limit of the sequence of rational numbers $\{x_i\}$ where the first $i-1$ decimal places of x_i agree with the first $i-1$ decimal places of $\sqrt{2}$, the i'th decimal place is equal to 3, and the rest of the decimal places are zero. Is $x = \sqrt{2}$? Give a reason for your answer.

15.7. Let x, y, and z be real numbers. Show the following properties hold.

(a) $x < y$ and $y < z$ implies $x < z$.

(b) $x < y$ implies $x + z < y + z$.

(c) $x < y$ implies $-x > -y$.

15.8. Find the set of x that satisfies (a) $|\sqrt{2}x - 3| \le 7$ and (b) $|3x - 6\sqrt{2}| > 2$.

15.9. Verify that the triangle inequality (7.14) extends to real numbers s and t.

15.10. *(Harder)* (a) If p is a rational number, x is a real number, and $\{x_i\}$ is any sequence of rational numbers that converges to x, show that $p < x$ implies that $p < x_i$ for all i sufficiently large. (b) If x and y are real numbers and $\{y_i\}$ is any sequence that converges to y, show that $x < y$ implies $x < y_i$ for all i sufficiently large.

15.11. Show that the following sequences are Cauchy sequences.

(a) $\left\{\frac{1}{(i+1)^2}\right\}$ (b) $\left\{4 - \frac{1}{2^i}\right\}$ (c) $\left\{\frac{i}{3i+1}\right\}$

15.12. Show that the sequence $\{i^2\}$ is **not** a Cauchy sequence.

15.13. Let $\{x_i\}$ denote the sequence of real numbers defined by

$$\begin{aligned}
x_1 &= .373373337\cdots \\
x_2 &= .337733377333377\cdots \\
x_3 &= .333777333377773333777\cdots \\
x_4 &= .3333777733333777773333337777\cdots \\
&\vdots
\end{aligned}$$

(a) Show that the sequence is a Cauchy sequence and (b) find $\lim_{i\to\infty} x_i$. This shows that a sequence of irrational numbers can converge to a rational number.

15.14. Can a number of the form $sx + t$, with s and t rational and x irrational, be rational?

15.15. Let $\{x_i\}$ and $\{y_i\}$ be Cauchy sequences with limits x and y respectively. (a) Show that $\{x_i - y_i\}$ is a Cauchy sequence and compute its limit. (b) Assuming there is a constant c such that $y_i \geq c > 0$ for all i, show that $\left\{\dfrac{x_i}{y_i}\right\}$ is a Cauchy sequence and compute its limit.

15.16. Show that a sequence that converges is a Cauchy sequence. Hint: if $\{x_i\}$ converges to x, write $x_i - x_j = (x_i - x) + (x - x_j)$ and use the triangle inequality.

15.17. *(Harder)* Let $\{x_i\}$ be an increasing sequence, $x_{i-1} \leq x_i$, which is bounded above, i.e. there is a number c such that $x_i \leq c$ for all i. Prove that $\{x_i\}$ converges. Hint: Use a variation of the argument for the convergence of the bisection algorithm

15.18. Compute the first 5 terms of the sequence that defines the value of the function $f(x) = \dfrac{x}{x+2}$ at $x = \sqrt{2}$. Hint: follow Fig. 15.5 and use the *evalf* function of *MAPLE*© in order to determine all the digits.

15.19. Let $\{x_i\}$ be the sequence with $x_i = 3 - \dfrac{2}{i}$ and $f(x) = x^2 - x$. What is the limit of the sequence $\{f(x_i)\}$?

15.20. Show that $|x|$ is Lipschitz continuous on the real numbers $\mathbb{R}$.

15.21. Let n be a natural number. Show that $\dfrac{1}{x^n}$ is Lipschitz continuous on the set of rational numbers $Q = \{x : .01 \leq x \leq 1\}$ and find a Lipschitz constant without using Theorem 15.3. Hint: Use the identity

$$\begin{aligned} x_2^n - x_1^n &= (x_2 - x_1)\big(x_2^{n-1} + x_2^{n-2}x_1 + x_2^{n-3}x_1^2 \\ &\qquad\qquad + \cdots + x_2^2 x_1^{n-3} + x_2 x_1^{n-2} + x_1^{n-1}\big) \\ &= (x_2 - x_1)\sum_{j=0}^{n-1} x_2^{n-1-j}\, x_1^j \end{aligned}$$

after showing that it is true. Note there are n terms in the last sum, the Lipschitz constant definitely depends on n.

15.22. Show Theorem 15.3 is true.

15.23. Write each of the following sets using the interval notation and then mark the sets on a number line.

(a) $\{x : -2 < x \leq 4\}$ (b) $\{x : -3 < x < -1\} \cup \{x : -1 < x \leq 2\}$
(c) $\{x : x = -2,\ 0 \leq x\}$ (b) $\{x : x < 0\} \cup \{x : x > 1\}$

15.24. Produce an interval that contains all the points $3 - 10^{-j}$ for $j \geq 0$ but does not contain 3.

15.25. Using *MATLAB*© or *MAPLE*© , graph the following functions on one graph: $y = 1 \times x$, $y = 1.4 \times x$, $y = 1.41 \times x$, $y = 1.414 \times x$, $y = 1.4142 \times x$, $y = 1.41421 \times x$. Use your results to explain how you could graph the function $y = \sqrt{2} \times x$.

15.26. (a) Give a definition of an interval (a, b) where a and b are real numbers in terms of intervals with rational endpoints. (b) Do the same for $[a, b]$.

15.27. Explain why there are infinitely many real numbers between any two distinct real numbers by giving a systematic way to write them down. Hint: first consider the case when the two distinct numbers are integers and work one digit at a time.

15.28. Find the Lipschitz constant of the function $f(x) = \sqrt{x}$ with $D(f) = (\delta, \infty)$ for given $\delta > 0$.

> The aim of Book X of Euclid's treatise on the "Elements" is to investigate the commensurable and the incommensurable, the rational and irrational continuous quantities. This science has its origin in the school of Pythagoras, but underwent an important development in the hands of the Athenian, Theaetetus, who is justly admired for his natural aptitude in this as in other branches of mathematics. One of the most gifted of men, he patiently pursued the investigation of truth contained in these branches of science ... and was in my opinion the chief means of establishing exact distinctions and irrefutable proofs with respect to the above mentioned quantities. (Pappus 290-350 (about))

16

The Bisection algorithm for $f(x) = 0$

Divide ut regnes (divide and conquer). (Machiavelli 1469-1527)

16.1 Bisection

We now generalize the Bisection algorithm used above to compute the positive root of the equation $x^2 - 2 = 0$, to compute roots of the equation

$$f(x) = 0 \tag{16.1}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a Lipschitz continuous function. The Bisection algorithm reads as follows, where TOL is a given positive tolerance:

1. Choose initial values x_0 and X_0 with $x_0 < X_0$ so that $f(x_0)f(X_0) < 0$. Set $i = 1$.
2. Given two rational numbers $x_{i-1} < X_{i-1}$ with the property that $f(x_{i-1})f(X_{i-1}) < 0$, set $\bar{x}_i = (x_{i-1} + X_{i-1})/2$.
   - If $f(\bar{x}_i) = 0$, then stop.
   - If $f(\bar{x}_i)f(X_{i-1}) < 0$, then set $x_i = \bar{x}_i$ and $X_i = X_i$.
   - If $f(\bar{x}_i)f(x_{i-1}) < 0$, then set $x_i = x_i$ and $X_i = \bar{x}_i$.
3. Stop if $X_i - x_i \leq TOL$.
4. Increase i by 1 and go back to step 2.

The equation $f(x) = 0$ may have many roots, and the choice of initial approximations x_0 and X_0 such that $f(x_0)f(X_0) \leq 0$ restricts the search for one or more roots to the interval $[x_0, X_0]$. To find all roots of an equation $f(x)$ it may be necessary to systematically search for all the possible start intervals $[x_0, X_0]$.

The proof that the Bisection algorithm converges is the same as that given above in the special case when $f(x) = x^2 - 0$. By construction, we have after i steps, assuming that we don't stop because $f(\bar{x}_i) = 0$ and $X_0 - x_0 = 1$, that

$$0 \leq X_i - x_i \leq 2^{-i},$$

and as before that

$$|x_i - x_j| \leq 2^{-i} \quad \text{if } j \geq i.$$

Again, $\{x_i\}_{i=1}^{\infty}$ is a Cauchy sequence and thus converges to a unique real number $\bar{x}$, and by construction

$$|x_i - \bar{x}| \leq 2^{-i} \quad \text{and} \quad |X_i - \bar{x}| \leq 2^{-i}.$$

It remains to show that $\bar{x}$ is a root of $f(x) = 0$, that is, we have to show that $f(\bar{x}) = 0$. By definition $f(\bar{x}) = f(\lim_{i\to\infty} x_i) = \lim_{i\to\infty} f(x_i)$ and thus we need to show that $\lim_{i\to\infty} f(x_i) = 0$. To this end we use the Lipschitz continuity to see that

$$|f(x_i) - f(X_i)| \leq L|x_i - X_i| \leq L2^{-i}.$$

Since $f(x_i)f(X_i) < 0$, that is the signs of $f(x_i)$ and $f(X_i)$ are different, this proves that in fact

$$|f(x_i)| \leq L2^{-i} \quad (\text{and also } |f(X_i)| \leq L2^{-i}),$$

which proves that $\lim_{i\to\infty} f(x_i) = 0$, and thus $f(\bar{x}) = \lim_{i\to\infty} f(x_i) = 0$ as we wanted to show.

We summarize this as a theorem, which is known as *Bolzano's Theorem* after the Catholic priest B. Bolzano (1781-1848), who was one of the first people to work out analytic proofs of properties of continuous functions.

Theorem 16.1 (Bolzano's Theorem) *If $f : [a, b] \to \mathbb{R}$ is Lipschitz continuous and $f(a)f(b) < 0$, then there is a real number $\bar{x} \in [a, b]$ such that $f(\bar{x}) = 0$.*

One consequence of this theorem is called the *Intermediate Value Theorem*, which states that if $g(x)$ is Lipschitz continuous on an interval $[a, b]$ then $g(x)$ takes on every value between $g(a)$ and $g(b)$ at least once as x varies over $[a, b]$. This follows applying Bolzanos theorem to the function $f(x) = g(x) - y = 0$, where y lies between $f(a)$ and $f(b)$.

Theorem 16.2 (The Intermediate Value Theorem) *If $f : [a, b] \to \mathbb{R}$ is Lipschitz continuous then for any real number y in the interval between $f(a)$ and $f(b)$, there is a real number $x \in [a, b]$ such that $f(x) = y$.*

FIGURE 16.1. Bernard Placidus Johann Nepomuk Bolzano 1781-1848, Czech mathematician, philosopher and catholic priest: "My special pleasure in mathematics rested therefore particularly on its purely speculative parts, in other words I prized only that part of mathematics which was at the same time philosophy".

16.2 An example

As an application of the Bisection algorithm, we compute the roots of the chemical equilibrium equation (7.13) in Chapter *Rational Numbers*,

$$S\,(.02 + 2S)^2 - 1.57 \times 10^{-9} = 0. \tag{16.2}$$

We show a plot of the function involved in Fig. 16.2. Apparently there are roots near $-.01$ and 0, but to compute them it seems advisable to first rescale the equation. We then first multiply both sides of (16.2) by 10^9 to get

$$10^9 \times S\,(.02 + 2S)^2 - 1.57 = 0,$$

and write

$$\begin{aligned} 10^9 \times\ S\,(.02 + 2S)^2 &= 10^3 \times S \times 10^6 \times (.02 + 2S)^2 \\ = 10^3 \times S \times \left(10^3\right)^2 \times (.02 + 2S)^2 &= 10^3 \times S \times \left(10^3 \times (.02 + 2S)\right)^2 \\ &= 10^3 \times S \times \left(20 + 2 \times 10^3 \times S\right)^2. \end{aligned}$$

If we name a new variable $x = 10^3 S$, then we obtain the following equation to solve

$$f(x) = x(20 - 2x)^2 - 1.57 = 0. \tag{16.3}$$

The polynomial $f(x)$ has more reasonable coefficients and the roots are not nearly as small as in the original formulation. If we find a root x of $f(x) = 0$, then we can find the physical variable $S = 10^{-3}x$. We note that

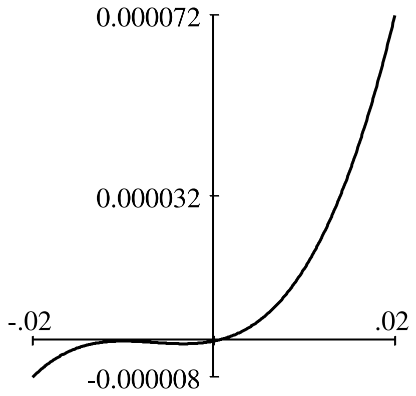


FIGURE 16.2. A plot of the function $S\,(.02 + 2S)^2 - 1.57 \times 10^{-9}$

only positive roots can have any meaning in this model, since we cannot have "negative" solubility.

The function $f(x)$ is a polynomial and thus is Lipschitz continuous on any bounded interval, and thus the Bisection algoritm can be used to compute its roots. We plot $f(x)$ in Fig. 16.3. It appears that $f(x) = 0$ might have

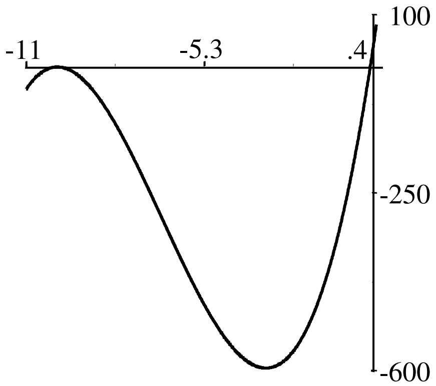


FIGURE 16.3. A plot of the function $f(x) = x(20 - 2x)^2 - 1.57$

one root near 0 and another root near -10.

To compute a positive root, we now choose $x_0 = -.1$ and $X_0 = .1$ and apply the Bisection algorithm for 20 steps. We show the results in Fig. 16.4. This suggests that the root of (16.3) is $x \approx .00392$ or $S \approx 3.92 \times 10^{-6}$.

i	x_i	X_i
0	-0.10000000000000	0.10000000000000
1	0.00000000000000	0.10000000000000
2	0.00000000000000	0.05000000000000
3	0.00000000000000	0.02500000000000
4	0.00000000000000	0.01250000000000
5	0.00000000000000	0.00625000000000
6	0.00312500000000	0.00625000000000
7	0.00312500000000	0.00468750000000
8	0.00390625000000	0.00468750000000
9	0.00390625000000	0.00429687500000
10	0.00390625000000	0.00410156250000
11	0.00390625000000	0.00400390625000
12	0.00390625000000	0.00395507812500
13	0.00390625000000	0.00393066406250
14	0.00391845703125	0.00393066406250
15	0.00391845703125	0.00392456054688
16	0.00392150878906	0.00392456054688
17	0.00392150878906	0.00392303466797
18	0.00392150878906	0.00392227172852
19	0.00392189025879	0.00392227172852
20	0.00392189025879	0.00392208099365

FIGURE 16.4. 20 steps of the Bisection algorithm applied to (16.3) using $x_0 = -.1$ and $X_0 = .1$.

16.3 Computational cost

We applied the Deca-section to compute $\sqrt{2}$ above. Of course we can use this method also for computing the root of a general equation. Once we have more than one method to compute a root of a equation, it is natural to ask which method is "best". We have to decide what we mean by "best" of course. For this problem, best might mean "most accurate" or "cheapest" for example. The exact criteria depends on our needs.

The criteria may depend on many things, such as the the level of accuracy to try to achieve. Of course, this depends on the application and the computational cost. In the Muddy Yard Model, a couple of decimal places is certainly sufficient from a practical point of view. If we actually tried to measure the diagonal using a tape measure for example, we would only get to within a few centimeters of the true value even neglecting the difficulty of measuring along a straight line. For more accuracy, we could use a laser and measure the distance to within a couple of wavelengths, and thus we might want to compute with a corresponding precision of many decimals. This would of course be overkill in the present case, but could be necessary in applications to e.g. astronomy or geodesic (for instance continental drift). In physics there is a strong need to compute certain quantities with many digits. For example one would like to know the mass of the electron very accurately. In applications of mechanics, a couple of decimals in the final answer may often be enough.

For the Deca-section and Bisection algorithms, accuracy is apparently not an issue, since both algorithms can be executed until we get 16 places or whatever number of digits is used for floating point representation. Therefore the way to compare the methods is by the amount of computing time it takes to achieve a given level of accuracy. This computing time is often called the *cost* of the computation, a left-over from the days when computer time was actually purchased by the second.

The cost involved in one of these algorithms can be determined by figuring out the cost per iteration step and then multiplying by the total number of steps we need to reach the desired accuracy. In one step of the Bisection Algorithm, the computer must compute the midpoint between two points, evaluate the function f at that point and store the value temporarily, check the sign of the function value, and then store the new x_i and X_i. We assume that the time it takes for the computer to do each of these operations can be measured and we define

$$\begin{aligned}
C_m &= \text{cost of computing the midpoint}\\
C_f &= \text{cost of evaluating } f \text{ at a point}\\
C_{\pm} &= \text{cost of checking the sign of a variable}\\
C_s &= \text{cost of storing a variable.}
\end{aligned}$$

The total cost of one step of the bisection algorithm is

$$C_m + C_f + C_\pm + 4C_s$$

and the cost after N_b steps is

$$N_b(C_m + C_f + C_\pm + 4C_s). \tag{16.4}$$

One step of the Deca-section algorithm has a considerably higher cost because there are 9 intermediate points to check. The total cost after N_d steps of the Deca-section algorithm is

$$N_d(9C_m + 9C_f + 9C_\pm + 20C_s). \tag{16.5}$$

On the other hand, the difference $|x_i - \bar{x}|$ decreases by a factor of $1/10$ after each step of the Deca-section algorithm as compared to a factor of $1/2$ after each step of the Bisection algorithm. Since $1/2^3 > 1/10 > 1/2^4$, this means that the Bisection algorithm requires between 3 and 4 times as many steps as the Deca-section algorithm in order to reduce the initial size $|x_0 - \bar{x}|$ by a given factor. So $N_b \approx 4N_d$. This gives the cost of the Bisection Algorithm as

$$4N_d(C_m + C_f + C_\pm + 4C_s) = N_d(4C_m + 4C_f + 4C_\pm + 16C_s)$$

as compared to (16.5). This means that the Bisection algorithm is cheaper to use than the Deca-section algorithm.

17

Do mathematicians quarrel?*

The proofs of Bolzano's and Weierstrass theorems have a decidedly non-constructive character. They do not provide a method for actually finding the location of a zero or the greatest or smallest value of a function with a prescribed degree of precision in a finite number of steps. Only the mere existence, or rather the absurdity of the non-existence, of the desired value is proved. This is another important instance where the "intuitionists" have raised objections; some have even insisted that such theorems be eliminated from mathematics. The student of mathematics should take this no more seriously than did most of the critics. (Courant)

I know that the great Hilbert said "We will not be driven out from the paradise Cantor has created for us", and I reply "I see no reason to walking in". (R. Hamming)

There is a concept which corrupts and upsets all others. I refer not to the Evil, whose limited realm is that of ethics; I refer to the infinite. (Borges).

Either mathematics is too big for the human mind or the human mind is more than a machine. (Gödel)

17.1 Introduction

Mathematics is often taught as an "absolute science" where there is a clear distinction between true and false or right and wrong, which should be universally accepted by all professional mathematicians and every enlightened layman. This is true to a large extent, but there are important aspects of mathematics where agreement has been lacking and still is lacking. The development of mathematics in fact includes as fierce quarrels as any other science. In the beginning of the 20th century, the very foundations of mathematics were under intense discussion. In parallel, a split between "pure" and "applied" mathematics developed, which had never existed before. Traditionally, mathematicians were generalists combining theoretical mathematical work with applications of mathematics and even work in mechanics, physics and other disciplines. Leibniz, Lagrange, Gauss, Poincare and von Neumann all worked with concrete problems from mechanics, physics and a variety of applications, as well as with theoretical mathematical questions.

In terms of the foundations of mathematics, there are different "mathematical schools" that view the basic concepts and axioms somewhat differently and that use somewhat different types of arguments in their work. The three principal schools are the *formalists*, the *logicists* and finally the *intuitionists*, also known as the *constructivists*.

As we explain below, we group both the formalists and the logicists together under an *idealistic* tradition and the the constructivists under a *realistic* tradition. It is possible to associate the idealistic tradition to an "aristocratic" standpoint and the realistic tradition to a "democratic" one. The history of the Western World can largely be be viewed as a battle between an idealistic/aristochratic and a realistic/democratic tradition. The Greek philosopher Plato is the portal figure of the idealistic/aristocratic tradition, while along with the scientific revolution initiated in the 16th century, the realistic/democratic tradition has taken a leading role in our society.

The debate between the formalists/logicists and the constructivists culminated in the 1930s, when the program put forward by the formalists/logicists suffered a strong blow from the logician Kurt Gödel. Gödel showed, to the surprise of world including great mathematicians like Hilbert, that in any axiomatic mathematical theory containing the axioms for the natural numbers, there are true facts which cannot be proved from the axioms. This is Gödel's famous *incompleteness theorem*.

Alan Turing (1912-54) took up a similar line of thought in the form of computability of real numbers in his famous 1936 article *On Computable Numbers, with an application to the Entscheidungsproblem*. In this paper Turing introduced an abstract machine, now called a *Turing machine*, which became the prototype of the modern programmable computer. Turing defined a computable number as real number whose decimal expansion could be produced by a Turing machine. He showed that π was computable, but claimed that most real numbers are not computable. He gave gave examples of "undecidable problems" formulated as the problem if the Turing machine would come to a halt or not, see Fig. 17.2.

FIGURE 17.1. Kurt Gödel (with Einstein 1950): "Every formal system is incomplete"

This was a clear defeat for the formalists/logicists and a corresponding victory for the constructivists. Paradoxically, soon after the defeat the formalists/logicists gained control of the mathematics departments and the constructivists left to create new departments of computer science and numerical analysis based on constructive mathematics. It appears that the trauma generated by Gödel's findings on the incompleteness of axiomatic methods was so strong that the earlier co-existence of the formalists/logicists and constructivists was no longer possible. Even today, the world of mathematics is heavily influenced by this split.

We will come back to the dispute between the formalists/logicists and constructivists below, and use it to illustrate fundamental aspects of mathematics which hopefully can help us to understand our subject better.

17.2 The formalists

The *formalist* school says that it does not matter what the basic concepts *actually* mean, because in mathematics we are just concerned with relations between the basic concepts whatever the meaning may be. Thus, we do not have to (and cannot) explain or define the basic concepts and can view mathematics as some kind of "game". However, a formalist would be very

FIGURE 17.2. Alan Turing: "I wonder if my machine will come to a halt?".

anxious to demonstrate that in his formal system it would not be possible to arrive at *contradictions*, in which case his game would be at risk of breaking down. A formalist would thus like to be absolutely sure about the *consistency* of his formal system. Further, a formalist would like to know that, at least in principle, he would be able to understand his own game fully, that is that he would in principle be able to give a mathematical explanation or proof of any true property of his game. The mathematician Hilbert was the leader of the formalist school. Hilbert was shocked by the results by Gödel.

17.3 The logicists and set theory

The logicists try to base mathematics on logic and *set theory*. Set theory was developed during the second half of the 19th century and the language of set theory has become a part of our every day language and is very much appreciated by both the formalist and logicist schools, while the constructivists have a more reserved attitude. A set is a collection of items, which are the elements of the set. An element of the set is said to belong to the set. For example, a dinner may be viewed as a set consisting of various dishes (entree, main course, dessert, coffee). A family (the Wilsons) may be viewed as a set consisting of a father (Mr. Wilson), a mother (Mrs. Wilson)

FIGURE 17.3. Bertrand Russell: "I am protesting"

and two kids (Tom and Mary). A soccer team (IFK Göteborg for example) consists of the set of players of the team. Humanity may be said to be set of all human beings.

Set theory makes it possible to speak about collections of objects as if they were single objects. This is very attractive in both science and politics, since it gives the possibility of forming new concepts and groups in hierarchical structures. Out of old sets, one may form new sets whose elements are the old sets. Mathematicians like to speak about *the set of all real numbers*, denoted by $\mathbb{R}$, *the set of all positive real numbers*, the *set of all prime numbers*, et cetera, and a politician planning a campaign may think of the set of democratic voters, the set of auto workers, the set of female first time voters, or the set of all poor, jobless, male criminals. Further, a workers union may be thought of as a set of workers in a particular factory or field, and workers unions may come together into unions or sets of workers unions.

A set may be described by listing all the elements of the set. This may be very demanding if the set contains many elements (for example if the set is humanity). An alternative is to describe the set through a property shared by all the elements of the set, e.g. the set of all people who have the properties of being poor, jobless, male, and criminal at the same time. To describe humanity as the set of beings which share the property of being human, however seems to more of a play with words than something very useful.

The leader of the logicist school was the philosopher and peace activist Bertrand Russell (1872-1970). Russell discovered that building sets freely can lead into contradictions that threaten the credibility of the whole logicist system. Russell created variants of the old *liars paradox* and *barbers* *paradox*, which we now recall. Gödel's theorem may be viewed to a variant of this paradox.

The liars paradox

The liars paradox goes as follows: A person says "I am lying". How should you interpret this sentence? If you assume that what the person says is indeed true, then it means that he is lying and then what he says is not true. On the other hand, if you assume that what he says is not true, this means that he is not lying and thus telling the truth, which means that what he says is true. In either case, you seem to be led to a contradiction, right? Compare Fig. 17.4.

FIGURE 17.4. "I am (not) lying"

The barbers paradox

The barbers paradox goes as follows: The barber in the village has decided to cut the hair of everyone in the village who does not cut his own hair. What shall the barber do himself? If he decides to cut his own hair, he will belong to the group of people who cut their own hair and then according to his decision, he should not cut his own hair, which leads to a contradiction. On the other hand, if he decides not to cut his own hair, then he would belong to the group of people not cutting their own hair and then according to his decision, he should cut his hair, which is again a contradiction. Compare Fig. 17.5.

FIGURE 17.5. Attitudes to the “barbers paradox”: one relaxed and one very concerned.

17.4 The constructivists

The *intuitionist/constructivist* view is to consider the basic concepts to have a meaning which may be directly ”intuitively” understood by our brains and bodies through experience, without any further explanation. Furthermore, the intuitionists would like to use as concrete or “constructive” arguments as possible, in order for their mathematics always to have an intuitive “real” meaning and not just be a formality like a game.

An intuitionist may say that the natural numbers 1, 2, 3,...., are obtained by repeatedly adding 1 starting at 1. We took this standpoint when introducing the natural numbers. We know that from the constructivist point of view, the natural numbers are something in the state of being created in a process without end. Given a natural number n, there is always a next natural number $n+1$ and the process never stops. A constructivist would not speak of the set of all natural numbers as something having been completed and constituting an entity in itself, like the set of all natural numbers as a formalist or logicist would be willing to do. Gauss pointed out that “the set of natural numbers” rather would reflect a “mode of speaking” than existence as a set.

An intuitionist would not feel a need of “justification” or a proof of consistency through some extra arguments, but would say that the justification is built into the very process of developing mathematics using constructive processes. A constructivist would so to speak build a machine that could fly (an airplane) and that very constructive process would itself be a proof of the claim that building an airplane would be possible. A constructivist is thus in spirit close to a practicing engineer. A formalist would not actually build an airplane, rather make some model of an airplane, and would then need some type of argument to convince investors and passengers that his

FIGURE 17.6. Luitzen Egbertus Jan Brouwer 1881-1966: :"One cannot inquire into the foundations and nature of mathematics without delving into the question of the operations by which mathematical activity of the mind is conducted. If one failed to take that into account, then one would be left studying only the language in which mathematics is represented rather than the essence of mathematics".

airplane would actually be able to fly, at least in principle. The leader of the intuitionist school was Brouwer (1881-1967),see Fig. 17.6. Hard-core constructivism makes life very difficult (like strong vegetarianism), and because the Brouwer school of constructivists were rather fundamentalist in their spirit, they were quickly marginalized and lost influence in the 1930s. The quote by Courant given above shows the strong feelings involved related to the fact that very fundamental dogmas were at stake, and the general lack of rational arguments to meet the criticism from the intuitionists, which was often replaced by ridicule and oppression.

Van der Waerden, mathematician who studied at Amsterdam from 1919 to 1923 wrote: "Brouwer came [to the university] to give his courses but lived in Laren. He came only once a week. In general that would have not been permitted - he should have lived in Amsterdam - but for him an exception was made. ... I once interrupted him during a lecture to ask a question. Before the next week's lesson, his assistant came to me to say that Brouwer did not want questions put to him in class. He just did not want them, he was always looking at the blackboard, never towards the students. ... Even though his most important research contributions were in topology, Brouwer never gave courses on topology, but always on - and only on - the foundations of intuitionism . It seemed that he was no longer convinved of his results in topology because they were not correct from the point of view of intuitionism, and he judged everything he had done before, his greatest output, false according to his philosophy. He was a very strange person, crazy in love with his philosophy".

17.5 The Peano axiom system for natural numbers

The Italian mathematician Peano (1858-1932) set up an axiom system for the natural numbers using as undefined concepts "natural number", "successor", "belong to", "set" and "equal to". His five axioms are

1. 1 is a natural number
2. 1 is not the successor of any other natural number
3. Each natural number n has a successor
4. If the successors of n and m are equal then so are n and m

There is a fifth axiom which is the axiom of *mathematical induction* stating that if a property holds for any natural number n, whenever it holds for the natural number preceding n and it holds for $n = 1$, then it holds for *all natural numbers*. Starting with these five axioms, one can derive all the basic properties of real numbers indicated above.

We see that the Peano axiom system tries to catch the essence of our intuitive feeling of natural numbers as resulting from successively adding 1 without ever stopping. The question is if we get a more clear idea of the natural numbers from the Peano axiom system than from our intuitive feeling. Maybe the Peano axiom system helps to identify the basic properties of natural numbers, but it is not so clear what the improved insight really consists of.

The logicist Russell proposed in *Principia Mathematica* to define the natural numbers using set theory and logic. For instance, the number 1 would be defined roughly speaking as the set of all singletons, the number two the set of all dyads or pairs, the number three as the set of all triples, et cetera. Again the question is if this adds insight to our conception of natural numbers?

17.6 Real numbers

Many textbooks in calculus start with the assumption that the reader is already familiar with *real numbers* and quickly introduce the notation $\mathbb{R}$ to denote the set of *all real numbers*. The reader is usually reminded that the real numbers may be represented as points on the *real line* depicted as a horizontal (thin straight black) line with marks indicating 1, 2, and maybe numbers like 1.1, 1.2, $\sqrt{2}$, π et cetera. This idea of basing *arithmetic*, that is numbers, on *geometry* goes back to Euclid, who took this route to get around the difficulties of irrational numbers discovered by the Pythagoreans. However, relying solely on arguments from geometry is very impractical and Descartes turned the picture around in the 17th century by basing geometry on arithmetic, which opened the way to the revolution of Calculus. The difficulties related to the evasive nature of irrational numbers encountered by the Pythagoreans, then of course reappeared, and the related questions concerning the very foundations of mathematics gradually developed into a quarrel with fierce participation of many of the greatest mathematicians which culminated in the 1930s, and which has shaped the mathematical world of today.

We have come to the standpoint above that a real number may be defined through its decimal expansion. A rational real number has a decimal expansion that eventually becomes periodic. An irrational real number has an expansion which is infinite and is not periodic. We have defined $\mathbb{R}$ as the set of all possible infinite decimal expansions, with the agreement that this definition is a bit vague because the meaning of "possible" is vague. We may say that we use a constructivist/intuitionist definition of $\mathbb{R}$.

The formalist/logicist would rather like to define $\mathbb{R}$ as the set of all infinite decimal expansions, or set of all Cauchy sequences of rational numbers, in what we called a universal Big Brother style above.

The set of real numbers is often referred to as the "continuum" of real numbers. The idea of a "continuum" is basic in classical mechanics where both space and time is supposed to be "continuous" rather than "discrete". On the other hand, in quantum mechanics, which is the modern version of mechanics on the scales of atoms and molecules, matter starts to show features of being discrete rather than continuous. This reflects the famous particle-wave duality in quantum mechanics with the particle being discrete and the wave being continuous. Depending on what glasses we use, phenomena may appear to be more or less discrete or continuous and no single mode of description seems to suffice. The discussions on the nature of real numbers may be rooted in this dilemma, which may never be resolved.

17.7 Cantor vs Kronecker

Let us give a glimpse of the discussion on the nature of real numbers through two of the key personalities, namely Cantor (1845-1918) in the *formalist* corner and Kronecker (1823-91), in the *constructivist* corner. These two mathematicians were during the late half of the 19th century involved in a bitter academic fight through their professional lives (which eventually led Cantor into a tragic mental disorder). Cantor created *set theory* and in particular a theory about sets with *infinitely* many elements, such as the set of natural numbers or the set of real numbers. Cantors theory was criticized by Kronecker, and many others, who simply could not believe in Cantors mental constructions or consider them to be really interesting. Kronecker took a down-to-earth approach and said that only sets with finitely many elements can be properly understood by human brains ("God created the integers, all else is the work of man"). Alternatively, Kronecker said that only mathematical objects that can be "constructed" in a *finite* number of steps actually "exist", while Cantor allowed infinitely many steps in a "construction". Cantor would say that the set of *all natural numbers* that is the set with the elements $1, 2, 3, 4, ...$, would "exist" as an object in itself as *the set of all natural numbers* which could be grasped by human brains, while Kronecker would deny such a possibility and reserve it to a higher being. Of course, Kronecker did not claim that there are only finitely many natural numbers or that there is a largest natural number, but he would (following Aristotle) say that the existence of arbitrarily large natural numbers is like a "potential" rather than an actual reality.

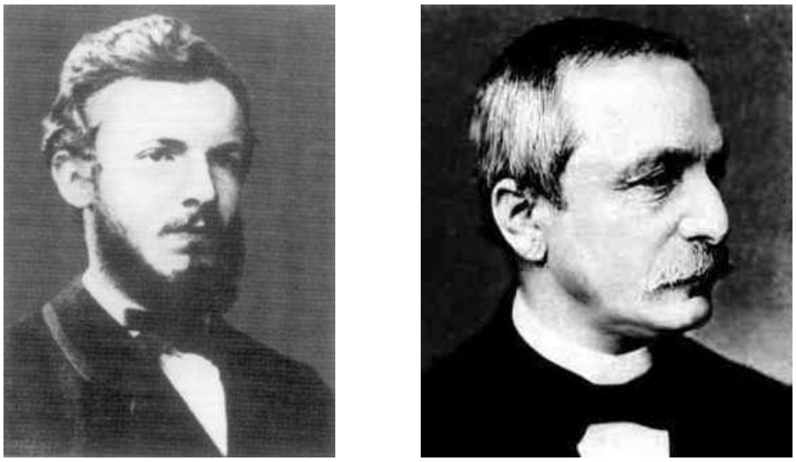


FIGURE 17.7. Cantor (left): "I realise that in this undertaking I place myself in a certain opposition to views widely held concerning the mathematical infinite and to opinions frequently defended on the nature of numbers". Kronecker (right): "God created the integers, all else is the work of man".

In the first round, Kronecker won since Cantor's theories about the infinite was rejected by many mathematicians in the late 19th and beginning 20th century. But in the next round, the influential mathematician Hilbert, the leader of the formalist school, joined on the side of Cantor. Bertrand Russell and Norbert Whitehead tried to give mathematics a foundation based on logic and set theory in their monumental *Principia Mathematica* (1910-13) and may also be viewed as supporters of Cantor. Thus, despite the strong blow from Gödel in the 1930's, the formalist/logicist schools took over the scene and have been dominating mathematics education into our time. Today, the development of the computer as is again starting to shift the weight to the side of the constructivists, simply because no computer is able to perform infinitely many operations nor store infinitely many numbers, and so the old battle may come alive again.

Cantor's theories about infinite numbers have mostly been forgotten, but there is one reminiscence in most presentations of the basics of Calculus, namely Cantors's argument that the degree of infinity of the real numbers is strictly larger than that of the rational or natural numbers. Cantor argued as follows: suppose we try to enumerate the real numbers in a list with a first real number r_1, a second real number r_2 and so on. Cantor claimed that in any such list there must be some real numbers missing, for example any real number that differs from r_1 in the first decimal, from r_2 in the second decimal and so on. Right? Kronecker would argue against this construction simply by asking full information about for example r_1, that is, full information about all the digits of r_1. OK, if r_1 was rational then this could be given, but if r_1 was irrational, then the mere listing of all the decimals of r_1 would never come to an end, and so the idea of a list of real numbers would not be very convincing. So what do you think? Cantor or Kronecker?

Cantor not only speculated about different degrees of infinities, but also cleared out more concrete questions about e.g. convergence of trigonometric series viewing real numbers as limits of of Cauchy sequences of rational numbers in pretty much the same we have presented.

17.8 Deciding whether a number is rational or irrational

We dwell a bit more on the nature of real numbers. Suppose x is a real number, the decimals of which can be determined one by one by using a certain algorithm. How can we tell if x is rational or irrational? Theoretically, if the decimal expansion is periodic then x is rational otherwise it is irrational. There is a practical problem with this answer however because we can only compute a finite number of digits, say never more than 10^{100}. How can we be sure that the decimal expansion does not start repeating after that? To be honest, this question seems very difficult to answer. Indeed it appears to be impossible to tell what happens in the complete decimal expansion by looking at a finite number of decimals. The only way to decide if a number x is rational or irrational is figure out a clever argument like the one the Pythagoreans used to show that $\sqrt{2}$ is irrational. Figuring out such arguments for different specific numbers like π and e is an activity that has interested a lot of mathematicians over the years.

On the other hand, the computer can only compute rational numbers and moreover only rational numbers with finite decimal expansions. If irrational numbers do not exist in practical computations, it is reasonable to wonder if they truly exist. Constructive mathematicians like Kronecker and Brouwer would not claim that irrational numbers really exist.

17.9 The set of all possible books

We suggest it is reasonable to define the set of all real numbers $\mathbb{R}$ as *the set of all possible decimal expansions* or equivalently *the set of all possible Cauchy sequences of rational numbers.* Periodic decimal expansions correspond to rational numbers and non-periodic expansions to irrational numbers. The set $\mathbb{R}$ thus consists of the set of all rational numbers together with the set of all irrational numbers. We know that it is common to omit the word "possible" in the suggested definition of $\mathbb{R}$ and define $\mathbb{R}$ as "the set of all real numbers", or "the set of all infinite decimal expansions".

Let's see if this hides some tricky point by way of an analogy. Suppose we define a "book" to be any finite sequence of letters. There are specific books such as "The Old Man and the Sea" by Hemingway, "The Author as a Young Dog" by Thomas, "Alice in Wonderland" by Lewis Carrol, and "1984" by Orwell, that we could talk about. We could then introduce **B** as "the set of all possible books", which would consist of all the books that have been and will be written purposely, together with many more "books" that consist of random sequences of letters. These would include those famous books that are written or could be written by chimpanzees playing with typewriters. We could probably handle this kind of terminology without too much difficulty, and we would agree that 1984 is an element of **B**. More generally, we would be able to say that any given book is a member of **B**. Although this statement is difficult to deny, it is also hard to say that this ability is very useful.

Suppose now we omit the word possible and start to speak of **B** as "the set of all books". This could give the impression that in some sense **B** is an existing reality, rather than some kind of potential as when we speak about "possible books". The set **B** could then be viewed as a library containing all books. This library would have to be enormously large and most of the "books" would be of no interest to anyone. Believing that the set of all books "exists" as a reality would not be very natural for most people.

The set of real numbers $\mathbb{R}$ has the same flavor as the set of all books **B**. It must be a very large set of numbers of which only a relative few, such as the rational numbers and a few specific irrational numbers, are ever encountered in practice. Yet, it is traditional to define $\mathbb{R}$ as the set of real numbers, rather than as "set of all possible real numbers". The reader may choose the interpretation of $\mathbb{R}$ according to his own taste. A true idealist would claim that the set of all real numbers "exists", while a down-to-earth person would more likely speak about the set of possible real numbers. Eventually, this may come down a personal religious feeling; some people appear to believe that Heaven actually exists, and while others might view as a potential or as a poetic way of describing something which is difficult to grasp.

Whatever interpretation you choose, you will certainly agree that some real numbers are more clearly specified than others, and that to specify a real number, you need to give some algorithm allowing you to determine as many digits of the real number as would be possible (or reasonable) to ask for.

17.10 Recipes and good food

Using the Bisection algorithm, we can compute any number of decimals of $\sqrt{2}$ if we have enough computational power. Using an algorithm to specify a number is analogous to using a recipe to specify for example *Grandpa's Chocolate Cake*. By following the recipe, we can bake a cake that is a more or less accurate approximation of the ideal cake (which only Grandpa can make) depending on our skill, energy, equipment and ingredients. There is a clear difference between the recipe and cakes made from the recipe, since after all we can eat a cake with pleasure but not a recipe. The recipe is like an algorithm or scheme telling us how to proceed, how many eggs to use for example, while cakes are the result of actually applying the algorithm with real eggs.

Of course, there are people who seem to enjoy reading recipes, or even just looking at pictures of food in magazines and talking about it. But if they never actually do cook anything, their friends are likely to lose interest in this activity. Similarly, you may enjoy looking at the symbols π, $\sqrt{2}$ et cetera, and talking about them, or writing them on pieces of paper, but if you you never actually compute them, you may come to wonder what you are actually doing.

In this book, we will see that there are many mathematical quantities that can only be determined approximately using a computational algorithm. Examples of such quantities are $\sqrt{2}$, π, and the base e of the natural logarithm. Later we will find that there are also functions, even elementary functions like $\sin(x)$ and $\exp(x)$ that need to be computed for different values of x. Just as we first need to bake a cake in order to enjoy it, we may need to compute such ideal mathematical quantities using certain algorithms before using them for other purposes.

17.11 The "new math" in elementary education

After the defeat of formalists in the 1930s by the arguments of Gödel, paradoxically the formalist school took over and set theory got a new chance. A wave generated by this development struck the elementary mathematics education in the 1960s in the form of the "new math". The idea was to explain numbers using set theory, just as Russell and Whitehead had tried to do 60 years earlier in their *Principia*. Thus a kid would learn that a set consisting of one cow, two cups, a piece of chocolate and an orange, would have five elements. The idea was to explain the nature of the number 5 this way rather than counting to five on the fingers or pick out 5 oranges from a heap of oranges. This type of "new math" confused the kids, and the parents and teachers even more, and was abandoned after some years of turbulence.

17.12 The search for rigor in mathematics

The formalists tried to give mathematics a rigorous basis. The search for rigor was started by Cauchy and Weierstrass who tried to give precise definitions of the concepts of limit, derivative and integral, and was continued by Cantor and Dedekind who tried to clarify the precise meaning of concepts such as continuum, real number, the set of real numbers et cetera. Eventually this effort of giving mathematics a fully rational basis collapsed, as we have indicated above.

We may identify two types of rigor:

- constructive rigor
- formal rigor.

Constructive rigor is necessary to accomplish difficult tasks like carrying out a heart operation, sending a man to the moon, building a tall suspension bridge, climbing Mount Everest, or writing a long computer program that works properly. In each case, every little detail may count and if the whole enterprise is not characterized by extreme rigor, it will most likely fail. Eventually this is a rigor that concerns material things, or real events.

Formal rigor is of a different nature and does not have a direct concrete objective like the ones suggested above. Formal rigor may be exercised at a royal court or in diplomacy, for example. It is a rigor that concerns language (words), or manners. The Scholastic philosophers during the Medieval time, were formalists who loved formal rigor and could discuss through very complicated arguments for example the question how many Angels could fit onto the edge of a knife. Some people use a very educated formally correct language which may be viewed as expressing a formal rigor. Authors pay a lot of attention to the formalities of language, and may spend hour after hour polishing on just one sentence until it gets just the right form. More generally, formal aspects may be very important in Arts and Aesthetics. Formal rigor may be thus very important, but serves a different purpose than constructive rigor. Constructive rigor is there to guarantee that something will actually function as desired. Formal rigor may serve the purpose of controlling people or impressing people, or just make people feel good, or to carry out a diplomatic negotiation. Formal rigor may be exercised in a game or play with certain very specific rules, that may be very strict, but do not serve a direct practical purpose outside the game.

Also in mathematics, one may distinguish between concrete and formal error. A computation, like multiplication of two natural numbers, is a concrete task and rigor simply means that the computation is carried out in a correct way. This may be very important in economics or engineering. It is not difficult to explain the usefulness of this type of constructive rigor, and the student has no difficulty in formulating himself what the criteria of constructive rigor might be in different contexts.

Formal rigor in calculus was promoted by Weierstrass with the objective of making basic concepts and arguments like the continuum of real numbers or limit processes more "formally correct". The idea of formal rigor is still alive very much in mathematics education dominated by the formalist school. Usually, students cannot understand the meaning of this type of "formally rigorous reasoning", and very seldom can exercise this type of rigor without much direction from the teacher.

We shall follow an approach where we try to reach constructive rigor to a degree which can be clearly motivated, and we shall seek to make the concept of formal rigor somewhat understandable and explain some of its virtues.

17.13 A non-constructive proof

We now give an example of a proof with non-constructive aspects that plays an important role in many Calculus books. Although because of the non-constructive aspects, the proof is considered to be so difficult that it can only by appreciated by selected math majors.

The setting is the following: We consider a bounded increasing sequence $\{a_n\}_1^\infty$ of real numbers, that is $a_n \leq a_{n+1}$ for $n = 1, 2, ...$, and there is s constant C such that $a_n \leq C$ for $n = 1, 2, ...$. The claim is that the sequence $\{a_n\}_1^\infty$ converges to a limit A. The proof goes as follows: all the numbers a_n clearly belong to the interval $I = [a_1, C]$. For simplicity suppose $a_1 = 0$ and $C = 1$. Divide now the interval $[0, 1]$ into the two intervals $[0, 1/2]$ and $[1/2, 1]$. and make the following choice: if there is a real number a_n such that $a_n \in [1/2, 1]$, then choose the right interval $[1/2, 1]$ and if not choose the left interval $[0, 1/2]$. Then repeat the subdivision into a left and a right interval, choose one of the intervals following the same principle: if there is a real number a_n in the right interval, then choose this interval, and if not choose the left interval. We then get a nested sequence of intervals with length tending to zero defining a unique real number that is easily seen to be the limit of the sequence $\{a_n\}_1^\infty$. Are you convinced? If not, you must be a constructivist.

So where is the hook of non-constructiveness in this proof? Of course, it concerns the choice of interval: in order to choose the correct interval you must be able to check if there is some a_n that belongs to the right interval, that is you must check if a_n belongs to the right interval for all sufficiently large n. The question from a constructivist point of view is if we can perform each check in a finite number of steps. Well, this may depend on the particular sequence $a_n \leq a_{n+1}$ under consideration. Let's first consider a sequence which is so simple that we may say that we know everything of interest: for example the sequence $\{a_n\}_1^\infty$ with $a_n = 1 - 2^{-n}$, that is the sequence $1/2, 3/4, 7/8, 15/16, 31/32, ...$, which is a bounded increasing sequence clearly converging to 1. For this sequence, we would be able to always choose the correct interval (the right one) because of its simplicity.

We now consider the sequence $\{a_n\}_1^\infty$ with $a_n = \sum_1^n \frac{1}{k^2}$, which is clearly an increasing sequence, and one can also quite easily show that the sequence is bounded. In this case the choice of interval is much more tricky, and it is not clear how to make the choice constructively without actually constructing the limit. So there we stand, and we may question the value of the non-constructive proof of existence of a limit, if we anyway have to construct the limit.

At any rate we sum up in the following result that we will use a a couple of times below.

Theorem 17.1 (non-constructive!) *A bounded increasing sequence converges.*

17.14 Summary

The viewpoint of Plato was to say that ideal points and lines exist in some Heaven above, while the points and lines which we as human beings can deal with, are some more or less incomplete copies or shades or images of the ideals. This is Plato's *idealistic* approach, which is related to the formalistic school. An intuitionist would say that we can never be sure of the existence of the ideals, and that we should concentrate on the more or less incomplete copies we can *construct* ourselves as human beings. The question of the actual existence of the ideals thus becomes a question of *metaphysics* or *religion*, to which there probably is no definite answer. Following our own feelings, we may choose to be either a idealist/formalist or an intuitionist/contructivist, or something in between.

The authors of this book have chosen such a middle way between the constructivist and formalist schools, trying always to be as constructive as is possible from a practical point of view, but often using a formalist language for reasons of convenience. The constructive approach puts emphasis on the concrete aspects of mathematics and brings it close to engineering and "body". This reduces the mystical character of mathematics and helps understanding. On the other hand, mathematics is not equal to engineering or only "body", and also the less concrete aspects or "soul" are useful for our thinking and in modeling the world around us. We thus seek a good synthesis of constructive and formalistic mathematics, or a synthesis of Body & Soul.

Going back to the start of our little discussion, we thus associate the logicist and formalistic schools with the idealistic/aristochratic tradition and the constructivists with the constructive/democratic tradition. As students, we would probably appreciate a constructive/democratic approach, since it aids the understanding and gives the student an active role. On the other hand, certain things indeed are very difficult to understand or construct, and then the idealistic/arisochratic approach opens a possible attitude to handle this dilemma.

The constructivist approach, whenever feasible, is appealing from educational point of view, since it gives the student an active role. The student is invited to construct himself, and not just watch an omnipotent teacher pick ready-made examples from Heaven.

Of course, the development of the modern computer has meant a tremendous boost of constructive mathematics, because what the computer does is constructive. Mathematics education is still dominated by the formalist school, and the most of the problems today afflicting mathematics education can be related to the over-emphasis of the idealistic school in times when constructive mathematics is dominating in applications.

Chapter 17 Problems

17.1. Can you figure out how the barber's paradox is constructed? Suppose the barber comes from another village. Does this resolve the paradox?

17.2. Another paradox of a similar kind goes as follows: Consider all the natural numbers which you can describe using at most 100 words or letters. For instance, you can describe the number 10 000 by the words "ten thousand" or "a one followed by four zeros". Describe now a number by specifying it as the smallest natural number which can not be described in at most one hundred words. But the sentence "the smallest natural number which can not be described in at most one hundred words" is a description of a certain number with fewer than 100 words (15 to be exact), which contradicts the very definition of the number as the number which could not be described with less than 100 words. Can you figure out how the paradox arises?

17.3. Describe as closely as you can what you mean by a *point* or *line*. Ask a friend to do the same, and try to figure out if your concepts are the same.

17.4. Study how the concept of real numbers is introduced by browsing through the first pages of some calculus books in your nearest library or on your book shelf.

17.5. Define the number $\omega \in (0,1)$ as follows: let the first digit of ω be equal to one if there are exactly 10 digits in a row equal to one in the decimal expansion of $\sqrt{2}$ and zero else, let the second be equal to one if there are exactly 20 digits in a row equal to one in the decimal expansion of $\sqrt{2}$ and zero else, and so on. Is ω a well defined real number? How many digits of ω could you think to be possible to compute?

17.6. Make a poll about what people think a real number is, from friends, relatives, politicians, rock musicians, to physics and mathematics professors.

> Some distinguished mathematicians have recently advocated the more or less complete banishment from mathematics of all non-constructive proofs. Even if such a program were desirable, it would involve tremendous complications and even the partial destruction of the body of living mathematics. For this reason it is no wonder that the school of "intuitionism", which has adopted this program, has met with strong resistance, and that even the most thoroughgoing intuitionists cannot always live up to their convictions. (Courant)

> The composition of vast books is a laborious and impoverishing extravagance. To go on for five hundred pages developing an idea whose perfect oral exposition is possible in a few minutes! A better course of procedure is to pretend that these books already exist, and then to offer a resume, a commentary...More reasonable, more inept, more indolent, I have preferred to write notes upon imaginary books. (Borges, 1941)

> I have always imagined that Paradise will be kind of a library. (Borges)

18
The function $y = x^r$

With equal passion I have sought knowledge. I have wished to understand the secrets of men. I have wished to know why the stars shine. And I have tried to apprehend the Pythagorean power by which numbers hold sway about the flux. A little of this, but not much, I have achieved. (Bertrand Russell 1872-1970).

18.1 The function $\sqrt{x}$

We showed above that we can solve the equation $x^2 = a$ for any positive rational number a using the Bisection algorithm. The unique positive solution is a real number denoted by $\sqrt{a}$. We can view $\sqrt{a}$ as a function of a defined for $a \in \mathbb{Q}_+$. Of course, we can extend the function $\sqrt{a}$ to $[0, \infty)$ since $0^2 = 0$ or $\sqrt{0} = 0$.

Changing names from a to x, we now consider the function $f(x) = \sqrt{x}$ with $D(f) = \mathbb{Q}_+$ and $f : \mathbb{Q}_+ \to \mathbb{R}_+$. As explained in the Chapter Real numbers, we can extend this into a function $f : \mathbb{R}_+ \to \mathbb{R}_+$ with $f(x) = \sqrt{x}$, using the Lipschitz continuity of $\sqrt{x}$ on intervals (δ, ∞) with $\delta > 0$ as discussed below. Since by definition $\sqrt{x}$ is the solution to the equation $y^2 = x$ with y as unknown, we have for $x \in \mathbb{R}^+$,

$$(\sqrt{x})^2 = x. \tag{18.1}$$

We plot the function $\sqrt{x}$ in Fig. 18.1.

The function $y = \sqrt{x}$ is increasing: if $x > \bar{x}$, then $\sqrt{x} > \sqrt{\bar{x}}$. Further, if $\{x_i\}$ is a sequence of positive real numbers with $\lim_{i\to\infty} x_i = 0$, then

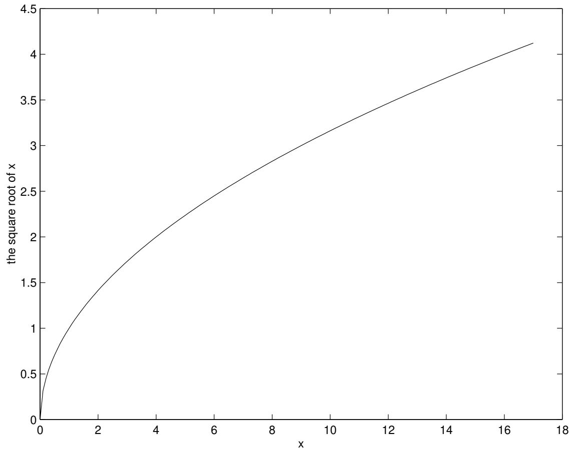


FIGURE 18.1. The function $\sqrt{x}$ of x.

obviously $\lim_{i\to\infty} \sqrt{x_i} = 0$, that is

$$\lim_{x\to 0^+} \sqrt{x} = 0. \tag{18.2}$$

18.2 Computing with the function $\sqrt{x}$

If $x^2 = a$ and $y^2 = b$, then $(xy)^2 = ab$. This gives the following property of the square root function,

$$\sqrt{a}\sqrt{b} = \sqrt{ab}. \tag{18.3}$$

We find similarly that

$$\frac{\sqrt{a}}{\sqrt{b}} = \sqrt{\frac{a}{b}}. \tag{18.4}$$

18.3 Is $\sqrt{x}$ Lipschitz continuous on $\mathbb{R}^+$?

To check if the function $f(x) = \sqrt{x}$ is Lipschitz continuous on $\mathbb{R}_+$, we note that since $(\sqrt{x} - \sqrt{\bar{x}})(\sqrt{x} + \sqrt{\bar{x}}) = x - \bar{x}$, we have

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

Since

$$\frac{1}{\sqrt{x} + \sqrt{\bar{x}}},$$

can be arbitrarily large by making x and $\bar{x}$ small positive, the function $f(x) = \sqrt{x}$ does not have a bounded Lipschitz constant on $\mathbb{R}_+$ and $f(x) = \sqrt{x}$ is *not* Lipschitz continuous on $\mathbb{R}_+$. This reflects the observation that the "slope" of $\sqrt{x}$ seems to increase without bound as x approaches zero. However, $f(x) = \sqrt{x}$ is Lipschitz continuous on any interval (δ, ∞) where δ is a fixed positive number, since we may then choose the Lipschitz constant L_f equal to $\frac{1}{2\delta}$.

18.4 The function x^r, where $r = \frac{p}{q}$ is rational

Consider the equation $y^q = x^p$ in the unknown y, where p and q are given integers and x is a given positive real number. Using the Bisection algorithm, we can prove that this equation has a unique solution y for any given positive x. We call the solution $y = x^{\frac{p}{q}} = x^r$, where $r = \frac{p}{q}$. In this way, we define a function $f(x) = x^r$ on $\mathbb{R}^+$ known as "x to the power r". Uniqueness follows from realizing that $y = x^r$ is increasing with x. Apparently, $\sqrt{x} = x^{\frac{1}{2}}$.

18.5 Computing with the function x^r

Using the defining equation $y^q = x^p$ as above, we find that for $x \in \mathbb{R}_+$ and $r, s \in \mathbb{Q}$,

$$x^r x^s = x^{r+s}, \quad \frac{x^r}{x^s} = x^{r-s}. \tag{18.5}$$

18.6 Generalizing the concept of Lipschitz continuity

There is a natural generalization of the concept of Lipschitz continuity that goes as follows. Let $0 < \theta \leq 1$ be a given number and L a positive constant, and suppose the function $f : \mathbb{R} \to \mathbb{R}$ satisfies

$$|f(x) - f(y)| \leq L|x - y|^\theta \quad \text{for all } x, y \in \mathbb{R}.$$

We say that $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous with exponent θ and Lipschitz constant L.

This generalizes the previous notion of Lipschitz continuity that corresponds to $\theta = 1$. Since θ can be smaller than one, we thus consider a larger class of functions. For example, we show that the function $f(x) = \sqrt{x}$ is Lipschitz continuous on $(0, \infty)$ with exponent $\theta = 1/2$ and Lipschitz constant $L = 1$, that is

$$|\sqrt{x} - \sqrt{\bar{x}}| \leq |x - \bar{x}|^{1/2}. \tag{18.6}$$

To prove this estimate, we assume that $x > \bar{x}$ and compute backwards, starting with $\bar{x} \leq \sqrt{x}\sqrt{\bar{x}}$ to get $x + \bar{x} - 2\sqrt{x}\sqrt{\bar{x}} \leq x - \bar{x} = |x - \bar{x}|$, which can be written

$$(\sqrt{x} - \sqrt{\bar{x}})^2 \leq (|x - \bar{x}|^{1/2})^2$$

from which the desired estimate follows by taking the square root. The case $\bar{x} > x$ is the same.

Functions that are Lipschitz continuous with Lipschitz exponent $\theta < 1$ may be quite "wild". In the "worst case", they may behave "everywhere" as "badly" as $\sqrt{x}$ does at $x = 0$. An example is given by Weierstrass function presented in the Chapter Fourier series. Take a look!

Chapter 18 Problems

18.1. Let $x,\ y \in \mathbb{R}$ and $r,\ s \in \mathbb{Q}$. Verify the following computing rules: (a) $x^{r+s} = x^r x^s$ (b) $x^{r-s} = x^r/x^s$ (c) $x^{rs} = (x^r)^s$ (d) $(xy)^r = x^r y^s$

18.2. Is $f(x) = \sqrt[3]{x}$, Lipschitz continuous on $(0, \infty)$ in the generalized sense? If yes give then the Lipschitz constant and exponent.

18.3. A Lipschitz continuous function with a Lipschitz constant L with $0 \leq L < 1$ is also called a *contraction mapping*. Which of the following functions are contraction mappings on $\mathbb{R}$? (a) $f(x) = \sin x$ (b) $f(x) = \frac{1}{1+x^2}$ (c) $f(x) = (1 + x^2)^{-1/2}$ (d) $f(x) = x^3$

18.4. Let $f(x) = 1$, for $x \leq 0$, and $f(x) = \sqrt{1 + x^2}$, for $x > 0$. Is f a contraction mapping?

18.5. Prove that a polynomial of degree greater than one cannot be a contraction mapping on $\mathbb{R}$.

18.6. Let $f : \mathbb{R} \to \mathbb{R}$ be defined as follows: $f(x) = 1/2$, if x is rational and $f(x) = 0$, if x is irrational. Is f a contraction mapping?

19
Fixed points and contraction mapping

Give me one fixed point on which to stand, and I will move the Earth. (Archimedes)

19.1 Introduction

A special case of the basic problem of solving an algebraic equation $f(x) = 0$ takes the form: find $\bar{x}$ such that

$$\bar{x} = g(\bar{x}), \qquad (19.1)$$

where $g : \mathbb{R} \to \mathbb{R}$ is a given Lipschitz continuous function. The equation (19.1) says that $\bar{x}$ is a *fixed point* of the function $y = g(x)$, that is the output value $g(\bar{x})$ is the same as the input value $\bar{x}$. Graphically, we seek the intersection of the graphs of the line $y = x$ and the curve $y = g(x)$, see Fig. 19.1.

To solve the equation $x = g(x)$, we could rewrite it as $f(x) = 0$ with (for example) $f(x) = x - g(x)$ and then apply the Bisection (or Decasection) algorithm to $f(x) = 0$. Note that the two equations $f(x) = 0$ with $f(x) = x - g(x)$ and $x = g(x)$ have exactly the same solutions, that is the two equations are *equivalent*.

In this chapter we consider a different algorithm for solving the equation (19.1) that is of central importance in mathematics. This is the *Fixed Point Iteration* algorithm, which takes the following form: Starting with some x_0,

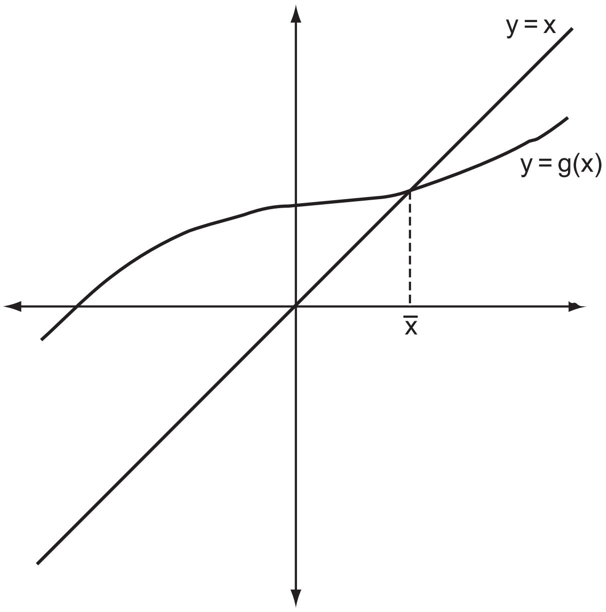


FIGURE 19.1. Illustration of a fixed point problem $g(\bar{x}) = \bar{x}$.

for $i = 1, 2, ...,$ compute

$$x_i = g(x_{i-1}) \quad \text{for } i = 1, 2, 3, ... \tag{19.2}$$

In words, we start with an initial approximation x_0 then compute $x_1 = g(x_0)$, $x_2 = g(x_1)$, $x_3 = g(x_2)$, and so on. Stepwise, given a current value x_{i-1}, we compute the corresponding output $g(x_{i-1})$, and then choose as new input $x_i = g(x_{i-1})$. Repeating this procedure, we will generate a sequence $\{x_i\}_{i=1}^{\infty}$.

We shall below study the following basic questions related to the sequence $\{x_i\}_{i=1}^{\infty}$ generated by Fixed Point Iteration:

- Does $\{x_i\}_{i=1}^{\infty}$ converge, that is does $\bar{x} = \lim_{i\to\infty} x_i$ exist?
- Is $\bar{x} = \lim_{i\to\infty} x_i$ a fixed point of $y = g(x)$, that is $\bar{x} = g(\bar{x})$?

We shall also investigate whether or not a fixed point $\bar{x}$ is uniquely determined.

19.2 Contraction mappings

We shall prove in this chapter that both the above questions have affirmative answers if $g(x)$ is Lipschitz continuous with Lipschitz constant $L < 1$, i.e.

$$|g(x) - g(y)| \leq L|x - y| \quad \text{for all } x, y \in \mathbb{R}, \tag{19.3}$$

with $L < 1$. We shall also see that the smaller L is, the quicker the convergence of the sequence $\{x_i\}$ to a fixed point, and the happier we will be.

A function $g : \mathbb{R} \to \mathbb{R}$ satisfying (19.3) with $L < 1$ is said to be a *contraction mapping*. We may summarize the basic result of this chapter as follows: A contraction mapping has a unique fixed point that is the limit of a sequence generated by Fixed Point Iteration. This is a most fundamental result of mathematics with a large number of applications. Sometimes it is referred to as Banach's contraction mapping theorem. Banach was a famous Polish mathematician, who created much of the field of *Functional Analysis*, which is a generalization of Calculus and Linear Algebra.

19.3 Rewriting $f(x) = 0$ as $x = g(x)$

Fixed Point Iteration is an algorithm for computing roots of equations of the form $x = g(x)$. If we are given an equation of the form $f(x) = 0$, we may want to rewrite this equation in the form of a fixed point equation $x = g(x)$. This can be done in many ways, for example by setting

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

where α is a nonzero real number to be chosen. Clearly, we have $\bar{x} = g(\bar{x})$ if and only if $f(\bar{x}) = 0$. To obtain quick convergence, one would try to choose α so that the Lipschitz constant of the corresponding function $g(x)$ is small. We shall see that trying to find such values of α leads to the wonderful world of *Newton methods* for solving equations, which is a very important part of mathematics.

A preliminary computation to find a good value of α to make $g(x) = x + \alpha f(x)$ have a small Lipschitz constant could go as follows. Assuming $x > y$,

$$\begin{aligned} g(x) - g(y) &= x + \alpha f(x) - (y + \alpha f(y)) = x - y + \alpha(f(x) - f(y)) \\ &= (1 + \alpha \frac{f(x) - f(y)}{x - y})|x - y|, \end{aligned}$$

which suggests choosing α to satisfy

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

We arrive at the same formula for $x < y$. We will return to this formula below. We note in particular the appearance of the quotient

$$\frac{f(x) - f(y)}{x - y},$$

which represents the slope of the *corda* or *secant* connecting the points $(x, f(x))$ and $(y, f(y))$ in $\mathbb{R}^2$, see Fig. 19.2

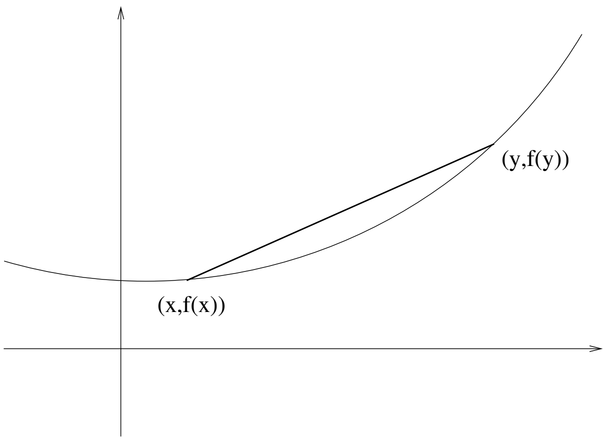


FIGURE 19.2. Corda connecting the points $(x, f(x))$ and $(y, f(y))$ in $\mathbb{R}^2$.

We now consider two models from everyday life leading to fixed point problems and apply the Fixed Point Iteration to solve them. In each case, the fixed point represents a balance or break-even of income and spending, with input equal to output. We then prove the contraction mapping theorem.

19.4 Card Sales Model

A door-to-door salesman selling greeting cards has a franchise with a greeting card company with the following price arrangement. For each shipment of cards, she pays a flat delivery fee of \$25 dollars and on top of this for sales of x, where x is measured in units of a hundred dollars, she pays an additional fee of 25% to the company. In mathematical terms, for sales of x hundreds of dollars, she pays

$$g(x) = \frac{1}{4} + \frac{1}{4}x \qquad (19.4)$$

where g is also given in units of a hundred dollars. The problem is to find the "break-even point", i.e. the amount of sales $\bar{x}$ where the money that she takes in $(= \bar{x})$ exactly balances the money she has to pay out $(g(\bar{x}))$, that is, her problem is to find the fixed point $\bar{x}$ satisfying $\bar{x} = g(\bar{x})$. Of course, she hopes to see that she clears a profit with each additional sale after this point.

We display the problem graphically in Fig. 19.3 in terms of two lines. The first line $y = x$ represents the amount of money collected for sales of x. In this problem, we measure sales in units of dollars, rather than say in numbers of cards sold, so we just get $y = x$ for this curve. The second line $y = g(x) = \frac{1}{4}x + \frac{1}{4}$ represents the amount of money that has to be paid to the greeting card company. Because of the initial flat fee of \$25, the salesman starts with a loss. Then as sales increase, she reaches the break-even point $\bar{x}$ and finally begins to see a profit.

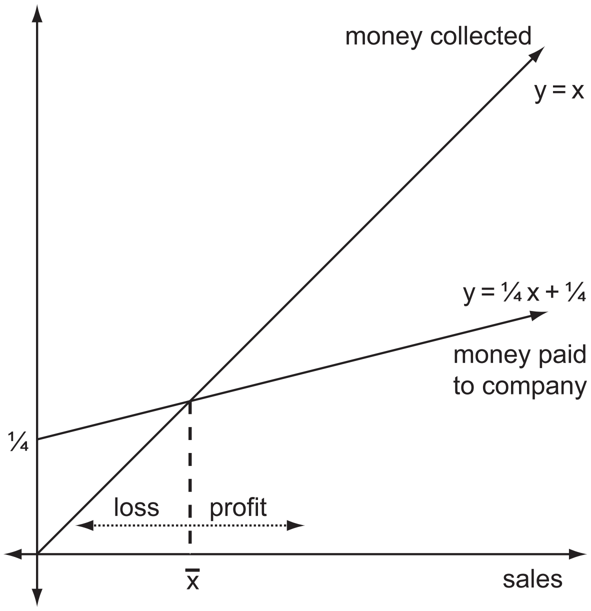


FIGURE 19.3. Illustration of the problem of determining the break-even point for selling greeting cards door-to-door. Sales above the break-even point $\bar{x}$ give a profit to the salesman, but sales below this point mean a loss.

In this problem, it is easy to analytically compute the break-even point, that is, the fixed point $\bar{x}$ because we can solve the equation

$$\bar{x} = g(\bar{x}) = \frac{1}{4}\bar{x} + \frac{1}{4}$$

to get $\bar{x} = 1/3$.

19.5 Private Economy Model

Your roommate has formulated the following model for her/his private economy: denote the net income by x that is variable including contributions from family, fellowship and a temporary job at McDonalds. The spending consists a fixed amount of 1 unit (of say 500 dollars per month) for rent and insurance, the variable amount of $x/2$ units for good food, good books and intellectual movies, and a variable amount of $1/x$ units for junk food, cigarettes and bad movies. This model is based on the observation that the more money your roommate has, the more educated a life she/he will live. The total spending is thus

$$g(x) = \frac{x}{2} + 1 + \frac{1}{x}$$

and the pertinent question is to find a balance of income and spending, that is to find the income $\bar{x}$ such that $\bar{x} = g(\bar{x})$ where the spending is the same as the income. If the income is bigger than $\bar{x}$, then your roommate will not use up all the money, which is against her/his nature, and if the income is less than $\bar{x}$, then your roommate's father will get upset, because he will have to pay the resulting debt.

Also in this case, we can directly find the fixed point $\bar{x}$ by solving the equation

$$\bar{x} = \frac{\bar{x}}{2} + 1 + \frac{1}{\bar{x}}$$

analytically and we then find that $\bar{x} = 1 + \sqrt{3} \approx 2.73$.

If we don't have enough motivation to go through the details of this calculation, we could instead try the Fixed Point Iteration. We would then start with an income $x_0 = 1$ say and compute the spending $g(1) = 2.5$, then choose the new income $x_1 = 2.5$, and compute the spending $g(x_1) = g(2.5) = 2.65$, and then set $x_2 = 2.65$ and compute the spending $g(x_2) = ...$ and so on. Of course, we expect that $\lim_i x_i = \bar{x} = 1 + \sqrt{3}$. Below, we will prove that this is indeed true!

19.6 Fixed Point Iteration in the Card Sales model

We now apply Fixed Point Iteration to the Card Sales Model. In Fig. 19.4, we plot the function $g(x) = \frac{1}{4}x + \frac{1}{4}$ along with $y = x$ and the fixed point $\bar{x}$. We also plot the value of $x_1 = g(x_0)$ for some initial approximation x_0. We choose $x_0 < \bar{x}$ because the sales start at zero and then increase. From the plot, we can see that $x_1 = g(x_0)$ is closer to $\bar{x}$ than x_0, i.e.

$$|g(x_0) - \bar{x}| < |x_0 - \bar{x}|.$$

In fact, we can compute the difference exactly since $\bar{x} = 1/3$,

$$|g(x_0) - \bar{x}| = \left|\frac{1}{4}x_0 + \frac{1}{4} - \frac{1}{3}\right| = \left|\frac{1}{4}\left(x_0 - \frac{1}{3}\right)\right| = \frac{1}{4}|x_0 - \bar{x}|.$$

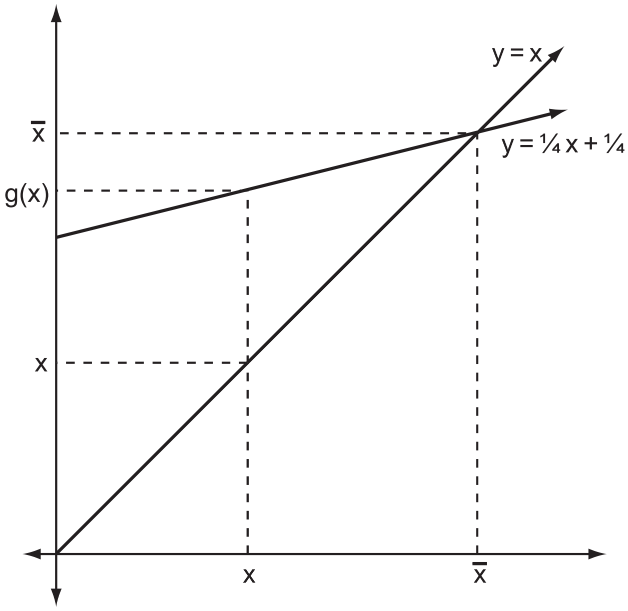


FIGURE 19.4. The first step of Fixed Point Iteration in Card Sales model: $g(x)$ is closer to $\bar{x}$ than x.

So the distance from $x_1 = g(x_0)$ to $\bar{x}$ is exactly 1/4 times the distance from x_0 to $\bar{x}$. The same argument shows that the distance from $x_2 = g(x_1)$ to $\bar{x}$ will be 1/4 of the distance from x_1 to $\bar{x}$ and thus 1/16 of the distance from x_0 to $\bar{x}$. In other words,

$$|x_2 - \bar{x}| = \frac{1}{4}|x_1 - \bar{x}| = \frac{1}{16}|x_0 - \bar{x}|$$

We illustrate this in Fig. 19.5. Generally, we have

$$|x_i - \bar{x}| = \frac{1}{4}|x_{i-1} - \bar{x}|,$$

and thus for $i = 1, 2, ...,$

$$|x_i - \bar{x}| = 4^{-i}|x_0 - \bar{x}|.$$

Since 4^{-i} gets as small as we please if i is sufficiently large, this estimate shows that Fixed Point Iteration applied to the Card Sales model converges, that is $\lim_{i\to\infty} x_i = \bar{x}$.

We consider some more examples before getting into the question of convergence of Fixed Point Iteration in a more general case.

EXAMPLE 19.1. For the sake of comparison, we show the results for the fixed point problem in Model 19.4 computed by applying the fixed

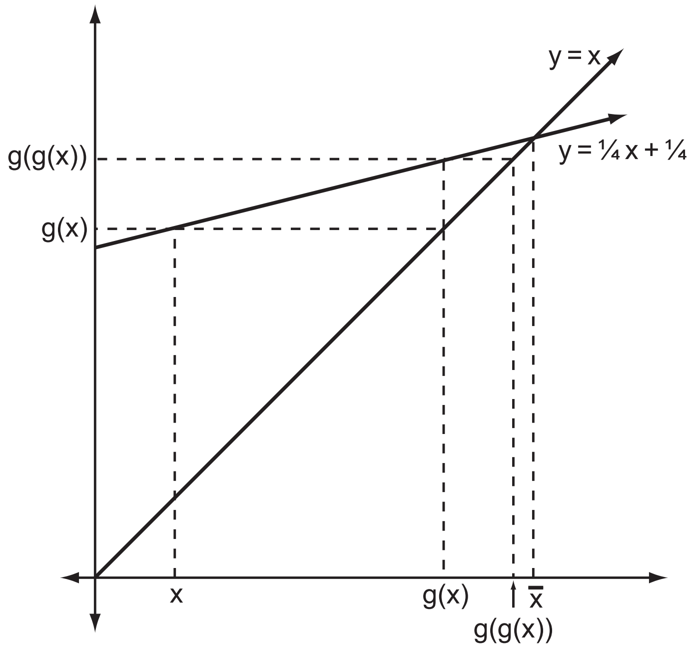


FIGURE 19.5. Two steps of the contraction map algorithm applied to the fixed point problem in Model 19.4. The distance of $g(g(x))$ to $\bar{x}$ is 1/4 the distance from $g(x)$ to $\bar{x}$ and 1/16 the distance from x to $\bar{x}$.

point iteration to $g(x) = \frac{1}{4}x + \frac{1}{4}$ and the bisection algorithm to the equivalent root problem for $f(x) = -\frac{3}{4}x + \frac{1}{4}$. To make the comparison fair, we use the initial value $x_0 = 1$ for the fixed point iteration and $x_0 = 0$ and $X_0 = 1$ for the bisection algorithm and compare the values of X_i from the bisection algorithm to x_i from the fixed point iteration in Fig. 19.6. The error of the fixed point iteration decreases by a factor

i	Bisection Algorithm X_i	Fixed Point Iteration x_i
0	1.00000000000000	1.00000000000000
1	0.50000000000000	0.50000000000000
2	0.50000000000000	0.37500000000000
3	0.37500000000000	0.34375000000000
4	0.37500000000000	0.33593750000000
5	0.34375000000000	0.33398437500000
6	0.34375000000000	0.33349609375000
7	0.33593750000000	0.33337402343750
8	0.33593750000000	
9	0.33398437500000	
10	0.33398437500000	
11	0.33349609375000	
12	0.33349609375000	
13	0.33337402343750	

FIGURE 19.6. Results of the bisection algorithm and the fixed point iteration used to solve the fixed point problem in Model 19.4. The error of the fixed point iteration decreases more for each iteration.

of 1/4 for each iteration as opposed to the error of the bisection algorithm which decreases by a factor of 1/2. This is clear in the table of results. Moreover, since both methods require one function evaluation and one storage per iteration but the bisection algorithm requires an additional sign check, the fixed point iteration costs less per iteration. We conclude that the fixed point iteration is truly "faster" than the bisection algorithm for this problem.

EXAMPLE 19.2. In solving for the solubility of $Ba(IO_3)_2$ in Model 7.10, we solved the root problem (16.3)

$$x(20 + 2x)^2 - 1.57 = 0$$

using the bisection algorithm. The results are in Fig. 16.4. In this example, we use the fixed point iteration to solve the equivalent fixed point problem

$$g(x) = \frac{1.57}{(20 + 2x)^2} = x. \tag{19.5}$$

We know that g is Lipschitz continuous on any interval that avoids $x = 10$ (and we also know that the fixed point/root is close to 0). We start off the iteration with $x_0 = 1$ and show the results in Fig. 19.7

i	x_i
0	1.00000000000000
1	0.00484567901235
2	0.00392880662465
3	0.00392808593169
4	0.00392808536527
5	0.00392808536483

FIGURE 19.7. Results of the fixed point iteration applied to (19.5).

EXAMPLE 19.3. In the case of the fixed point iteration applied to the Card Sales model, we can compute the iterates explicitly:

$$x_1 = \frac{1}{4}x_0 + \frac{1}{4}$$

and

$$\begin{aligned} x_2 &= \frac{1}{4}x_1 + \frac{1}{4} = \frac{1}{4}\left(\frac{1}{4}x_0 + \frac{1}{4}\right) + \frac{1}{4} \\ &= \frac{1}{4^2}x_0 + \frac{1}{4^2} + \frac{1}{4} \end{aligned}$$

Likewise, we find

$$x_3 = \frac{1}{4^3}x_0 + \frac{1}{4^3} + \frac{1}{4^2} + \frac{1}{4}$$

and after n steps

$$x_n = \frac{1}{4^n}x_0 + \sum_{i=1}^{n} \frac{1}{4^i}. \tag{19.6}$$

The first term on the right-hand side of (19.6), $\frac{1}{4^n}x_0$ converges to 0 as n increases to infinity. The second term is equal to

$$\sum_{i=1}^{n} \frac{1}{4^i} = \frac{1}{4} \times \sum_{i=0}^{n-1} \frac{1}{4^i} = \frac{1}{4} \times \frac{1 - \frac{1}{4^n}}{1 - \frac{1}{4}} = \frac{1 - \frac{1}{4^n}}{3}$$

using the formula for the geometric sum. The second term therefore converges to 1/3, which is precisely the fixed point for (19.4), as n increases to infinity.

An important observation about the last example is that the iteration converges because the slope of $g(x) = \frac{1}{4}x + \frac{1}{4}$ is $1/4 < 1$. This produces a factor of 1/4 for each iteration, forcing the right-hand side of (19.6) to have a limit as n tends to infinity. Recalling that the slope of a linear function is the same thing as its Lipschitz constant, we can say this example worked because the Lipschitz constant of g is $L = 1/4 < 1$.

In contrast if the Lipschitz constant, or slope, of g is larger than 1 then the analog of (19.6) will not converge. We demonstrate this graphically in Fig. 19.8 using the function $g(x) = 2x+\frac{1}{4}$. The difference between successive

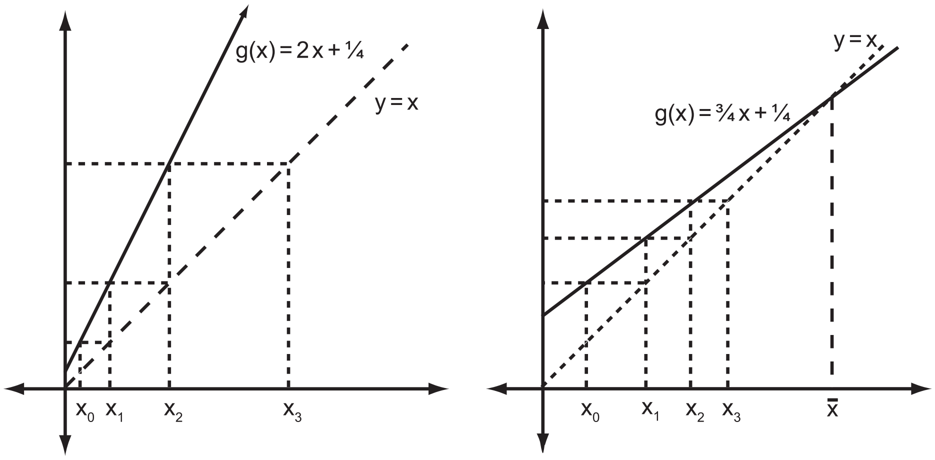


FIGURE 19.8. On the left, we plot the first three fixed point iterates for $g(x) = 2x + \frac{1}{4}$. The iterates increase without bound as the iteration proceeds. On the right, we plot the first three fixed point iterates for $g(x) = \frac{3}{4}x + \frac{1}{4}$. The iteration converges to the fixed point in this case.

iterates increases with each iteration and the fixed point iteration does not converge. It is clear from the plot that there is no positive fixed point. On the other hand, the fixed point iteration will converge when applied to any linear function with Lipschitz constant $L < 1$. We illustrate the convergence for $g(x) = \frac{3}{4}x + \frac{1}{4}$ in Fig. 19.8. Thinking about (19.6), the reason is simply that the geometric series with factor L converges when $L < 1$.

19.7 A contraction mapping has a unique fixed point

We now go back to the general case presented in the introductory overview. We shall prove that a contraction mapping $g : \mathbb{R} \to \mathbb{R}$ has a unique fixed point $\bar{x} \in \mathbb{R}$ given as the limit of a sequence generated by Fixed Point Iteration. We recall that a contraction mapping $g : \mathbb{R} \to \mathbb{R}$ is a Lipschitz continuous function on $\mathbb{R}$ with Lipschitz constant $L < 1$. We organize the proof as follows:

1. Proof that $\{x_i\}_{i=1}^{\infty}$ is a Cauchy sequence.
2. Proof that $\bar{x} = \lim_{i\to\infty} x_i$ is a fixed point.
3. Proof that $\bar{x}$ is unique.

Proof that $\{x_i\}_{i=1}^{\infty}$ is a Cauchy sequence

To estimate $|x_i - x_j|$ for $j > i$, we shall first prove an estimate for two consecutive indices, that is an estimate for $|x_{k+1} - x_k|$. To this end, we subtract the equation $x_k = g(x_{k-1})$ from $x_{k+1} = g(x_k)$ to get

$$x_{k+1} - x_k = g(x_k) - g(x_{k-1}).$$

Using the Lipschitz continuity of $g(x)$, we thus have

$$|x_{k+1} - x_k| \le L|x_k - x_{k-1}|. \tag{19.7}$$

Similarly,

$$|x_k - x_{k-1}| \le L|x_{k-1} - x_{k-2}|,$$

and thus

$$|x_{k+1} - x_k| \le L^2|x_{k-1} - x_{k-2}|.$$

Repeating the argument, we find that

$$|x_{k+1} - x_k| \le L^k|x_1 - x_0|. \tag{19.8}$$

We now proceed to use this estimate to estimate $|x_i - x_j|$ for $j > i$. We have

$$|x_i - x_j| = |x_i - x_{i+1} + x_{i+1} - x_{i+2} + x_{i+2} - \cdots + x_{j-1} - x_j|,$$

so that by the triangle inequality,

$$|x_i - x_j| \le |x_i - x_{i+1}| + |x_{i+1} - x_{i+2}| + \cdots + |x_{j-1} - x_j| = \sum_{k=i}^{j-1} |x_k - x_{k+1}|.$$

We now use (54.19) on each term $|x_k - x_{k+1}|$ in the sum to get

$$|x_i - x_j| \le \sum_{k=i}^{j-1} L^k\,|x_1 - x_0| = |x_1 - x_0| \sum_{k=i}^{j-1} L^k.$$

We compute

$$\sum_{k=i}^{j-1} L^k = L^i(1 + L + L^2 + \cdots + L^{j-i-1}) = L^i\,\frac{1 - L^{j-i}}{1 - L},$$

using the formula for the sum of a geometric series. We now use the assumption that $L < 1$, to conclude that $0 \leq 1 - L^{j-i} \leq 1$ and therefore for $j > i$,

$$|x_i - x_j| \leq \frac{L^i}{1-L}|x_1 - x_0|.$$

Since $L < 1$, the factor L^i can be made as small as we please by taking i large enough, and thus $\{x_i\}_{i=1}^{\infty}$ is a Cauchy sequence and therefore converges to a limit $\bar{x} = \lim_{i\to\infty} x_i$.

Note that the idea of estimating $|x_i - x_j|$ for $j > i$ by estimating $|x_k - x_{k+1}|$ and using the formula for a geometric sum is fundamental and will be used repeatedly below.

Proof that $\bar{x} = \lim_i x_i$ is a fixed point

Since $g(x)$ is Lipschitz continuous, we have

$$g(\bar{x}) = g(\lim_{i\to\infty} x_i) = \lim_{i\to\infty} g(x_i).$$

By the nature of the Fixed Point Iteration with $x_i = g(x_{i-1})$, we have

$$\lim_{i\to\infty} g(x_{i-1}) = \lim_{i\to\infty} x_i = \bar{x}.$$

Since of course

$$\lim_{i\to\infty} g(x_{i-1}) = \lim_{i\to\infty} g(x_i),$$

we thus see that $g(\bar{x}) = \bar{x}$ as desired. We conclude that the limit $\lim_i x_i = \bar{x}$ is a fixed point.

Proof of uniqueness

Suppose that x and y are two fixed points, that is $x = g(x)$ and $y = g(y)$. Since $g : \mathbb{R} \to \mathbb{R}$ is a contraction mapping,

$$|x - y| = |g(x) - g(y)| \leq L|x - y|$$

which is possible only if $x = y$ since $L < 1$. This completes the proof.

We have now proved that a contraction mapping $g : \mathbb{R} \to \mathbb{R}$ has a unique fixed point given by Fixed Point Iteration. We summarize in the following basic theorem.

Theorem 19.1 *A contraction mapping $g : \mathbb{R} \to \mathbb{R}$ has a unique fixed point $\bar{x} \in \mathbb{R}$, and any sequence $\{x_i\}_{i=1}^{\infty}$ generated by Fixed Point Iteration converges to $\bar{x}$.*

19.8 Generalization to $g : [a, b] \to [a, b]$

We may directly generalize this result by replacing $\mathbb{R}$ by any closed interval $[a, b]$ of $\mathbb{R}$. Taking the interval $[a, b]$ to be closed guarantees that $\lim_i x_i \in [a, b]$ if $x_i \in [a, b]$. It is critical that g maps the interval $[a, b]$ into *itself.*

Theorem 19.2 *A contraction mapping $g : [a, b] \to [a, b]$ has a unique fixed point $\bar{x} \in [a, b]$ and a sequence $\{x_i\}_{i=1}^{\infty}$ generated by Fixed Point Iteration starting with a point x_0 in $[a, b]$ converges to $\bar{x}$.*

EXAMPLE 19.4. We apply this theorem to $g(x) = x^4/(10-x)^2$. We can show that g is Lipschitz continuous on $[-1, 1]$ with $L = .053$ and the fixed point iteration started with any x_0 in $[-1, 1]$ converges rapidly to the fixed point $\bar{x} = 0$. However, the Lipschitz constant of g on $[-9.9, 9.9]$ is about 20×10^6 and the fixed point iteration diverges rapidly if $x_0 = 9.9$.

19.9 Linear convergence in Fixed Point Iteration

Let $\bar{x} = g(\bar{x})$ be the fixed point of a contraction mapping $g : \mathbb{R} \to \mathbb{R}$ and $\{x_i\}_{i=1}^{\infty}$ a sequence generated by Fixed Point Iteration. We can easily get an estimate on how quickly the error of the fixed point iterate x_i decreases as i increases, that is the speed of convergence, as follows. Since $\bar{x} = g(\bar{x})$, we have

$$|x_i - \bar{x}| = |g(x_{i-1}) - g(\bar{x})| \leq L|x_{i-1} - \bar{x}|, \tag{19.9}$$

which shows that the error decreases by *at least* a factor of $L < 1$ during each iteration. The smaller L is the faster the convergence!

The error may actually decrease by exactly a factor of L, as in the Card Sales model with $g(x) = \frac{1}{4}x + \frac{1}{4}$, where the error decreases by exactly a factor of $L = 1/4$ in each iteration.

When the error decreases by (at least) a constant factor $\theta < 1$ in each step, we say that the convergence is *linear* with *convergence factor* θ. The Fixed Point Iteration applied to a contraction mapping $g(x)$ with Lipschitz constant $L < 1$ converges linearly with convergence factor L.

We compare in Fig. 19.9. the speed of convergence of Fixed Point Iteration applied to $g(x) = \frac{1}{9}x + \frac{3}{4}$ and $g(x) = \frac{1}{5}x + 2$. The iteration for $\frac{1}{9}x + \frac{3}{4}$ reaches 15 places of accuracy within 15 iterations while the iteration for $\frac{1}{5}x + 2$ has only 14 places of accuracy after 20 iterations.

19.10 Quicker convergence

The functions $\frac{1}{2}x$ and $\frac{1}{2}x^2$ are both Lipschitz continuous on $[-1/2, 1/2]$ with Lipschitz constant $L = 1/2$, and have a unique fixed point $\bar{x} = 0$. The

i	x_i for $\frac{1}{9}x+\frac{3}{4}$	x_i for $\frac{1}{5}x+2$
0	1.00000000000000	1.00000000000000
1	0.86111111111111	2.20000000000000
2	0.84567901234568	2.44000000000000
3	0.84396433470508	2.48800000000000
4	0.84377381496723	2.49760000000000
5	0.84375264610747	2.49952000000000
6	0.84375029401194	2.49990400000000
7	0.84375003266799	2.49998080000000
8	0.84375000362978	2.49999616000000
9	0.84375000040331	2.49999923200000
10	0.84375000004481	2.49999984640000
11	0.84375000000498	2.49999996928000
12	0.84375000000055	2.49999999385600
13	0.84375000000006	2.49999999877120
14	0.84375000000001	2.49999999975424
15	0.84375000000000	2.49999999995085
16	0.84375000000000	2.49999999999017
17	0.84375000000000	2.49999999999803
18	0.84375000000000	2.49999999999961
19	0.84375000000000	2.49999999999992
20	0.84375000000000	2.49999999999998

FIGURE 19.9. Results of the fixed point iterations for $\frac{1}{9}x+\frac{3}{4}$ and $\frac{1}{5}x+2$.

estimate (19.9) suggests the fixed point iteration for both should converge to $\bar{x} = 0$ at the same rate. We show the results of the fixed point iteration applied to both in Fig. 19.10. We see that Fixed Point Iteration converges

i	x_i for $\frac{1}{2}x$	x_i for $\frac{1}{2}x^2$
0	0.50000000000000	0.50000000000000
1	0.25000000000000	0.25000000000000
2	0.12500000000000	0.06250000000000
3	0.06250000000000	0.00390625000000
4	0.03125000000000	0.00001525878906
5	0.01562500000000	0.00000000023283
6	0.00781250000000	0.00000000000000

FIGURE 19.10. Results of the fixed point iterations for $\frac{1}{2}x$ and $\frac{1}{2}x^2$.

much more quickly for $\frac{1}{2}x^2$, reaching 15 places of accuracy after 7 iterations. The estimate (19.9) thus does not give the full picture.

We now take a closer look into the argument behind (19.9) for the particular function $g(x) = \frac{1}{2}x^2$. As above we have with $\bar{x} = 0$,

$$x_i - 0 = \frac{1}{2}x_{i-1}^2 - \frac{1}{2}0^2 = \frac{1}{2}(x_{i-1} + 0)(x_{i-1} - 0),$$

and thus

$$|x_i - 0| = \frac{1}{2}|x_{i-1}|\,|x_{i-1} - 0|.$$

We conclude that the error of Fixed Point Iteration for $\frac{1}{2}x^2$ decreases by a factor of $\frac{1}{2}|x_{i-1}|$ during the i'th iteration. In other words,

for $i = 1$ the factor is $\frac{1}{2}|x_0|$,

for $i = 2$ the factor is $\frac{1}{2}|x_1|$,

for $i = 3$ the factor is $\frac{1}{2}|x_2|$,

and so on. We see that the reduction factor depends on the value of the current iterate.

Now consider what happens as the iteration proceeds and the iterates x_{i-1} become closer to zero. The factor by which the error in each step decreases becomes smaller as i increases! In other words, the closer the iterates get to zero, the faster they get close to zero. The estimate in (19.9) significantly *overestimates* the error of the fixed point iteration for $\frac{1}{2}x^2$ because it treats the error as if it decreases by a fixed factor each time. Thus it cannot be used to predict the rapid convergence for this function. For a function g, the first part of (19.9) tells the same story:

$$|x_i - \bar{x}| = |g(x_{i-1}) - g(\bar{x})|.$$

The error of x_i is determined by the change in g in going from $\bar{x}$ to the previous iterate x_{i-1}. This change can depend on x_{i-1} and when it does, the fixed point iteration does not converge linearly.

19.11 Quadratic convergence

We now consider a second basic example, where we establish quadratic convergence. We know that the Bisection algorithm for computing the root of $f(x) = x^2 - 2$ converges linearly with convergence factor $1/2$: the error gets reduced by the factor $\frac{1}{2}$ after each step. We can write the equation $x^2 - 2 = 0$ as the following fixed point equation

$$x = g(x) = \frac{1}{x} + \frac{x}{2}. \tag{19.10}$$

To see this, it suffices to multiply the equation (19.10) by x. We now apply Fixed Point Iteration to (19.10) to compute $\sqrt{2}$ and show the result in Fig. 19.11. We note that it only takes 5 iterations to reach 15 places of

i	x_i
0	1.00000000000000
1	1.50000000000000
2	1.41666666666667
3	1.41421568627451
4	1.41421356237469
5	1.41421356237310
6	1.41421356237310

FIGURE 19.11. The fixed point iteration for (19.10).

accuracy. The convergence appears to be very quick.

To see how quick the convergence in fact is, we seek a relation between the error in two consecutive steps. Computing as in (19.9), we find that

$$\begin{aligned} |x_i - \sqrt{2}| &= |g(x_{i-1}) - g(\sqrt{2})| \\ &= \left| \frac{x_{i-1}}{2} + \frac{1}{x_{i-1}} - \left(\frac{\sqrt{2}}{2} + \frac{1}{\sqrt{2}}\right) \right| \\ &= \left| \frac{x_{i-1}^2 + 2}{2x_{i-1}} - \sqrt{2} \right|. \end{aligned}$$

Now we find a common denominator for the fractions on the right and then use the fact that

$$(x_{i-1} - \sqrt{2})^2 = x_{i-1}^2 - 2\sqrt{2}x_{i-1} + 2$$

to get

$$|x_i - \sqrt{2}| = \frac{(x_{i-1} - \sqrt{2})^2}{2x_{i-1}} \approx \frac{1}{2\sqrt{2}}(x_{i-1} - \sqrt{2})^2. \tag{19.11}$$

We conclude that the error in x_i is the square of the error of x_{i-1} up to the factor $\frac{1}{2\sqrt{2}}$. This is *quadratic* convergence, which is very quick. In each step of the iteration, the number of correct decimals doubles!

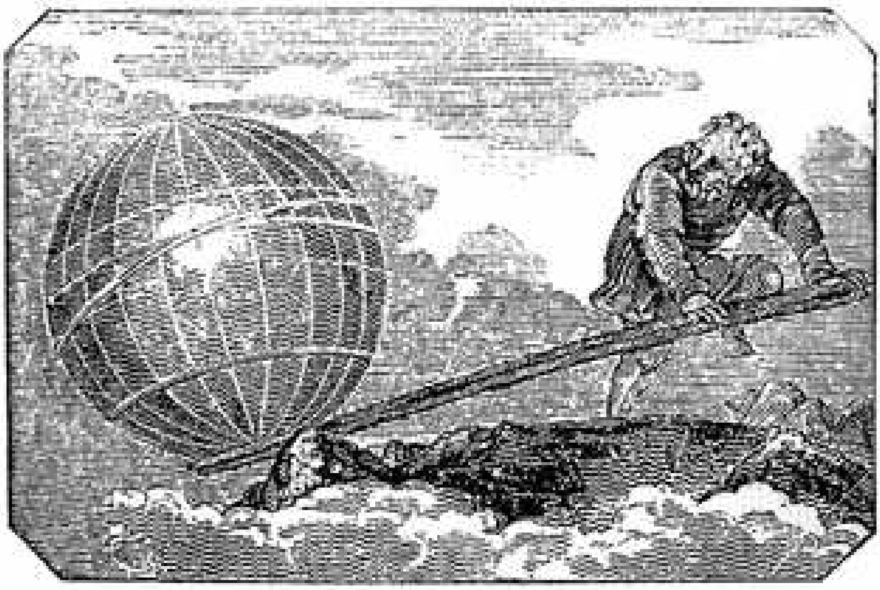


FIGURE 19.12. Archimedes moving the Earth with a lever and a fixed point.

Chapter 19 Problems

19.1. A salesman selling vacuum cleaners door-to-door has a franchise with the following payment scheme. For each delivery of vacuum cleaners, the salesman pays a fee of \$100 and then a percentage of the sales, measured in units of hundreds of dollars, that increases as the sales increases. For sales of x, the percentage is $20x\%$. Show that this model gives a fixed point problem and make a plot of the fixed point problem that shows the location of the fixed point.

19.2. Rewrite the following fixed point problems as root problems three different ways each.

$$\text{(a) } \frac{x^3-1}{x+2} = x \qquad \text{(b) } x^5 - x^3 + 4 = x$$

19.3. Rewrite the following root problems as fixed point problems three different ways each.

$$\text{(a) } 7x^5 - 4x^3 + 2 = 0 \qquad \text{(b) } x^3 - \frac{2}{x} = 0$$

19.4. (a) Draw a Lipschitz continuous function g on the interval $[0, 1]$ that has three fixed points such that $g(0) > 0$ and $g(1) < 1$. (b) Draw a Lipschitz continuous function g on the interval $[0, 1]$ that has three fixed points such that $g(0) > 0$ and $g(1) > 1$.

19.5. Write a program that implements Algorithm 19.2. The program should employ two methods for stopping the iteration: (1) when the number of iterations is larger than a user-input number and (2) when the difference between successive iterates $|x_i - x_{i-1}|$ is smaller than a user-input tolerance. Test the program by reproducing the results in Fig. 19.9 that were computed using *MATLAB*© .

19.6. In Section 7.10, suppose that K_{sp} for $Ba(IO_3)_2$ is 1.8×10^{-5}. Find the solubility S to 10 decimal places using the program from Proposition 19.5 after writing the problem as a suitable fixed point problem. Hint: $1.8 \times 10^{-5} = 18 \times 10^{-6}$ and $10^{-6} = 10^{-2} \times 10^{-4}$.

19.7. In Section 7.10, determine the solubility of $Ba(IO_3)_2$ in a .037 mole/liter solution of KIO_3 to 10 decimal places using the program from Proposition 19.5 after writing the problem as a suitable fixed point problem.

19.8. The power P delivered into a load R of a simple class A amplifier of output resistance Q and output voltage E is

$$P = \frac{E^2 R}{(Q+R)^2}.$$

Find all possible solutions R for $P = 1$, $Q = 3$, and $E = 4$ to 10 decimal places using the program from Proposition 19.5 after writing the problem as a fixed point problem.

19.9. Van der Waal's model for one mole of an ideal gas including the effects of the size of the molecules and the mutual attractive forces is

$$\left(P + \frac{a}{V^2}\right)(V - b) = RT,$$

where P is the pressure, V is the volume of the gas, T is the temperature, R is the ideal gas constant, a is a constant depending on the size of the molecules and the attractive forces, and b is a constant depending on the volume of all the molecules in one mole. Find all possible volumes V of the gas corresponding to $P = 2$, $T = 15$, $R = 3$, $a = 50$, and $b = .011$ to 10 decimal places using the program from Proposition 19.5 after writing the problem as a fixed point problem.

19.10. Verify that (19.6) is true.

19.11. (a) Find an explicit formula (similar to (19.6)) for the n'th fixed point iterate x_n for the function $g(x) = 2x + \frac{1}{4}$. (b) Prove that x_n diverges to ∞ as n increases to ∞.

19.12. (a) Find an explicit formula (similar to (19.6)) for the n'th fixed point iterate x_n for the function $g(x) = \frac{3}{4}x + \frac{1}{4}$. (b) Prove that x_n converges as n increases to ∞ and compute the limit.

19.13. (a) Find an explicit formula (similar to (19.6)) for the n'th fixed point iterate x_n for the function $g(x) = mx + b$. (b) Prove that x_n converges as n increases to ∞ provided that $L = |m| < 1$ and compute the limit.

19.14. Draw a Lipschitz continuous function g that does *not* have the property that x in $[0, 1]$ means that $g(x)$ is in $[0, 1]$.

19.15. (a) If possible, find intervals suitable for application of the fixed point iteration to each of the three fixed point problems found in Problem 19.3(a). (b) If possible, find intervals suitable for application of the fixed point iteration to each of the three fixed point problems found in Problem 19.3(b). In each case, a suitable interval is one on which the function is a contraction map.

19.16. *Harder* Apply Theorem 54.5 to the function $g(x) = 1/(1 + x^2)$ to show that the fixed point iteration converges on any interval $[a, b]$.

19.17. Given the following results of the fixed point iteration applied to a function $g(x)$,

i	x_i
0	14.00000000000000
1	14.25000000000000
2	14.46875000000000
3	14.66015625000000
4	14.82763671875000
5	14.97418212890625

compute the Lipschitz constant L for g. Hint: consider (54.19).

19.18. Verify the details of Example 19.4.

19.19. Given the following results of the fixed point iteration applied to a function $g(x)$,

i	x_i
0	0.50000000000000
1	0.70710678118655
2	0.84089641525371
3	0.91700404320467
4	0.95760328069857
5	0.97857206208770

decide if the convergence rate is linear or not.

19.20. (a) Show that $g(x) = \frac{2}{3}x^3$ is Lipschitz continuous on $[-1/2, 1/2]$ with Lipschitz constant $L = 1/2$. (b) Use the program from Problem 19.5 to compute 6 fixed point iterations starting with $x_0 = .5$ and compare to the results in Fig. 19.10. (c) Show that the error of x_i is approximately the cube of the error of x_{i-1} for any i.

19.21. Verify that (19.11) is true.

19.22. (a) Show the root problem $f(x) = x^2 + x - 6$ can be written as the fixed point problem $g(x) = x$ with $g(x) = \dfrac{6}{x+1}$. Show that the error of x_i decreases at a linear rate to the fixed point $\bar{x} = 2$ when the fixed point iteration converges to 2 and estimate the convergence factor for x_i close to 2. (b) Show the root problem $f(x) = x^2 + x - 6$ can be written as the fixed point problem $g(x) = x$ with $g(x) = \dfrac{x^2+6}{2x+1}$. Show that the error of x_i decreases at a quadratic rate to the fixed point $\bar{x} = 2$ when the fixed point iteration converges to 2.

19.23. The *Regula Falsi Method* is a variation of the bisection method for computing a root of $f(x) = 0$. For $i \geq 1$, assuming $f(x_{i-1})$ and $f(x_i)$ have the opposite signs, define x_{i+1} as the point where the straight line through $(x_{i-1}, f(x_{i-1}))$ and $(x_i, f(x_i))$ intersects the x-axis. Write this method as fixed point iteration by giving an appropriate $g(x)$ and estimate the corresponding convergence factor.

20
Analytic geometry in $\mathbb{R}^2$

> Philosophy is written in the great book (by which I mean the Universe) which stands always open to our view, but it cannot be understood unless one first learns how to comprehend the language and interpret the symbols in which it is written, and its symbols are triangles, circles, and other geometric figures, without which it is not humanly possible to comprehend even one word of it; without these one wanders in a dark labyrinth. (Galileo)

20.1 Introduction

We give a brief introduction to *analytic geometry* in two dimensions, that is the linear algebra of the *Euclidean plane*. Our common school experience has given us an intuitive *geometric* idea of the Euclidean plane as an infinite flat surface without borders consisting of points, and we also have an intuitive geometric idea of geometric objects like straight lines, triangles and circles in the plane. We brushed up our knowledge and intuition in geometry somewhat in Chapter *Pythagoras and Euclid*. We also presented the idea of using a coordinate system in the Euclidean plane consisting of two perpendicular copies of $\mathbb{Q}$, where each point in the plane has two coordinates (a_1, a_2) and we view $\mathbb{Q}^2$ as the set of ordered pairs of rational numbers. With only the rational numbers $\mathbb{Q}$ at our disposal, we quickly run into trouble because we cannot compute distances between points in $\mathbb{Q}^2$. For example, the distance between the points $(0, 0)$ and $(1, 1)$, the length of the diagonal of a unit square, is equal to $\sqrt{2}$, which is not a rational number. The troubles are resolved by using real numbers, that is by extending $\mathbb{Q}^2$ to $\mathbb{R}^2$.

In this chapter, we present basic aspects of analytic geometry in the Euclidean plane using a coordinate system identified with $\mathbb{R}^2$, following the fundamental idea of Descartes to describe geometry in terms of numbers. Below, we extend to analytic geometry in three-dimensional Euclidean space identified with $\mathbb{R}^3$ and we finally generalize to analytic geometry in $\mathbb{R}^n$, where the dimension n can be any natural number. Considering $\mathbb{R}^n$ with $n \geq 4$ leads to *linear algebra* with a wealth of applications outside Euclidean geometry, which we will meet below. The concepts and tools we develop in this chapter focussed on Euclidean geometry in $\mathbb{R}^2$ will be of fundamental use in the generalizations to geometry in $\mathbb{R}^3$ and $\mathbb{R}^n$ and linear algebra.

The tools of the geometry of Euclid is the ruler and the compasses, while the tool of analytic geometry is a calculator for computing with numbers. Thus we may say that Euclid represents a form of *analog* technique, while analytic geometry is a *digital* technique based on numbers. Today, the use of digital techniques is exploding in communication and music and all sorts of virtual reality.

20.2 Descartes, Inventor of Analytic Geometry

The foundation of modern science was laid by René Descartes (1596-1650) in *Discours de la méthod pour bien conduire sa raison et chercher la vérité dans les sciences* from 1637. *The Method* contained as an appendix *La Geometrie* with the first treatment of Analytic Geometry. Descartes believed that only mathematics may be certain, so all must be based on mathematics, the foundation of the *Cartesian* view of the World.

In 1649 Queen Christina of Sweden persuaded Descartes to go to Stockholm to teach her mathematics. However the Queen wanted to draw tangents at 5 a.m. and Descartes broke the habit of his lifetime of getting up at 11 o'clock, c.f. Fig. 20.1. After only a few months in the cold Northern climate, walking to the palace at 5 o'clock every morning, he died of pneumonia.

20.3 Descartes: Dualism of Body and Soul

Descartes set the standard for studies of Body and Soul for a long time with his *De homine* completed in 1633, where Descartes proposed a mechanism for automatic reaction in response to external events through nerve fibrils, see Fig. 20.2. In Descartes' conception, the rational Soul, an entity distinct from the Body and making contact with the body at the pineal gland,

FIGURE 20.1. Descartes: "The Principle which I have always observed in my studies and which I believe has helped me the most to gain what knowledge I have, has been never to spend beyond a few hours daily in thoughts which occupy the imagination, and a few hours yearly in those which occupy the understanding, and to give all the rest of my time to the relaxation of the senses and the repose of the mind".

might or might not become aware of the differential outflow of *animal spirits* brought about though the nerve fibrils. When such awareness did occur, the result was conscious sensation – Body affecting Soul. In turn, in voluntary action, the Soul might itself initiate a differential outflow of animal spirits. Soul, in other words, could also affect Body.

In 1649 Descartes completed *Les passions de l'ame*, with an account of causal Soul/Body interaction and the conjecture of the localization of the Soul's contact with the Body to the pineal gland. Descartes chose the pineal gland because it appeared to him to be the only organ in the brain that was not bilaterally duplicated and because he believed, erroneously, that it was uniquely human; Descartes considered animals as purely physical automata devoid of mental states.

20.4 The Euclidean plane $\mathbb{R}^2$

We choose a *coordinate system* for the Euclidean plane consisting of two straight lines intersecting at a 90° angle at a point referred to as the *origin*. One of the lines is called the x_1-axis and the other the x_2-axis, and each line is a copy of the real line $\mathbb{R}$. The *coordinates* of a given point a in the plane is the ordered pair of real numbers (a_1, a_2), where a_1 corresponds to the intersection of the x_1-axis with a line through a parallel to the x_2-axis, and a_2 corresponds to the intersection of the x_2-axis with a line through

FIGURE 20.2. Automatic reaction in response to external stimulation from Descartes *De homine* 1662

a parallel to the x_1-axis, see Fig. 20.3. The coordinates of the origin are $(0, 0)$.

In this way, we identify each point a in the plane with its coordinates (a_1, a_2), and we may thus represent the Euclidean plane as $\mathbb{R}^2$, where $\mathbb{R}^2$ is the set of ordered pairs (a_1, a_2) of real numbers a_1 and a_2. That is

$$\mathbb{R}^2 = \{(a_1, a_2) : a_1, a_2 \in \mathbb{R}\}.$$

We have already used $\mathbb{R}^2$ as a coordinate system above when plotting a function $f : \mathbb{R} \to \mathbb{R}$, where pairs of real numbers $(x, f(x))$ are represented as geometrical points in a Euclidean plane on a book-page.

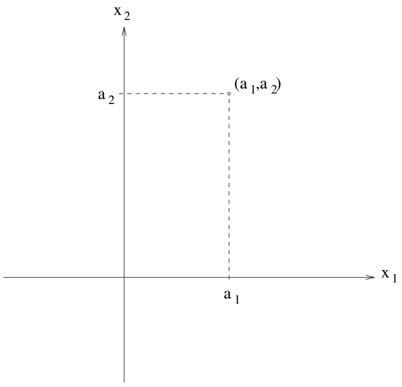


FIGURE 20.3. Coordinate system for $\mathbb{R}^2$

To be more precise, we can identify the Euclidean plane with $\mathbb{R}^2$, once we have chosen the (i) origin, and the (ii) direction (iii) scaling of the coordinate axes. There are many possible coordinate systems with different origins and orientations/scalings of the coordinate axes, and the coordinates of a geometrical point depend on the choice of coordinate system. The need to change coordinates from one system to another thus quickly arises, and will be an important topic below.

Often, we orient the axes so that the x_1-axis is horizontal and increasing to the right, and the x_2-axis is obtained rotating the x_1 axis by $90°$, or a quarter of a complete revolution counter-clockwise, see Fig. 20.3 or Fig. 20.4 displaying MATLAB's view of a coordinate system. The positive direction of each coordinate axis may be indicated by an arrow in the direction of increasing coordinates.

However, this is just one possibility. For example, to describe the position of points on a computer screen or a window on such a screen, it is not uncommon to use coordinate systems with the origin at the upper left corner and counting the a_2 coordinate positive down, negative up.

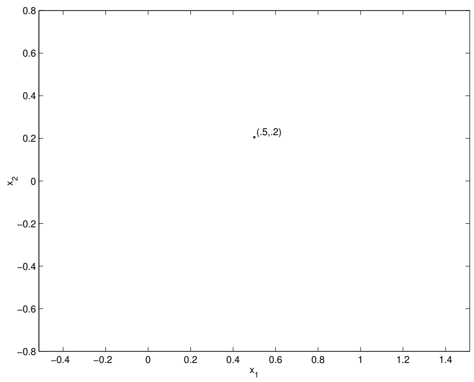


FIGURE 20.4. Matlabs way of visualizing a coordinate system for a plane.

20.5 Surveyors and Navigators

Recall our friends the Surveyor in charge of dividing land into properties, and the Navigator in charge of steering a ship. In both cases we assume that the distances involved are sufficiently small to make the curvature of the Earth negligible, so that we may view the world as $\mathbb{R}^2$. Basic problems faced by a Surveyor are (s1) to locate points in Nature with given coordinates on a map and (s2) to compute the area of a property knowing its corners. Basic problems of a Navigator are (n1) to find the coordinates on a map of his present position in Nature and (n2) to determine the present direction to follow to reach a point of destiny.

We know from Chapter 2 that problem (n1) may be solved using a GPS navigator, which gives the coordinates (a_1, a_2) of the current position of the GPS-navigator at a press of a button. Also problem (s1) may be solved using a GPS-navigator iteratively in an 'inverse" manner: press the button and check where we are and move appropriately if our coordinates are not the desired ones. In practice, the precision of the GPS-system determines its usefulness and increasing the precision normally opens a new area of application. The standard GPS with a precision of 10 meters may be OK for a navigator, but not for a surveyor, who would like to get down to meters or centimeters depending on the scale of the property. Scientists measuring continental drift or beginning landslides, use an advanced form of GPS with a precision of millimeters.

Having solved the problems (s1) and (n1) of finding the coordinates of a given point in Nature or vice versa, there are many related problems of type (s2) or (n2) that can be solved using mathematics, such as computing the area of pieces of land with given coordinates or computing the direction of a piece of a straight line with given start and end points. These are examples of basic problems of geometry, which we now approach to solve using tools of analytic geometry or linear algebra.

20.6 A first glimpse of vectors

Before entering into analytic geometry, we observe that $\mathbb{R}^2$, viewed as the set of ordered pairs of real numbers, can be used for other purposes than representing positions of geometric points. For example to describe the current weather, we could agree to write $(27, 1013)$ to describe that the temperature is 27 C$^\circ$ and the air pressure 1013 millibar. We then describe a certain weather situation as an ordered pair of numbers, such as $(27, 1013)$. Of course the *order* of the two numbers is critical for the interpretation. A weather situation described by the pair $(1013, 27)$ with temperature 1013 and pressure 27, is certainly very different from that described by $(27, 1013)$ with temperature 27 and pressure 1013.

Having liberated ourselves from the idea that a pair of numbers must represent the coordinates of a point in a Euclidean plane, there are endless possibilities of forming pairs of numbers with the numbers representing different things. Each new interpretation may be viewed as a new interpretation of $\mathbb{R}^2$.

In another example related to the weather, we could agree to write $(8, NNE)$ to describe that the current wind is 8 m/s and headed North-North-East (and coming from South-South-East. Now, NNE is not a real number, so in order to couple to $\mathbb{R}^2$, we replace NNE by the corresponding angle, that is by 22.5° counted positive clockwise starting from the North direction. We could thus indicate a particular wind speed and direction by the ordered pair $(8, 22.5)$.

You are no doubt familiar with the weather man's way of visualizing such a wind on the weather map using an arrow. In Fig. 20.5, we can see that the direction of the wind is NNE, and the length of the wind arrow is supposed to correspond to the wind speed 8 (using some appropriate scale).

FIGURE 20.5. The wind in Gothenburg is 8 (m/s) and headed NNE.

The wind arrow in Fig. 20.5 could also be described in terms of another pair of parameters, namely by how much it extends to the East and to the North respectively, that is by the pair $(8\sin(22.5°), 8\cos(22.5°)) \approx (3.06, 7.39)$. We could say that 3.06 is the "amount of East", and 7.39 is the "amount of North" of the wind velocity, while we may say that the wind *speed* is 8, where we think of the speed as the "absolute value" of the wind *velocity* $(3.06, 7.39)$. We thus think of the wind velocity as having both a direction, and an "absolute value" or "length". In this case, we view an ordered pair (a_1, a_2) as a *vector*, rather than as a point, and we can then represent the vector by an arrow.

We will soon see that ordered pairs viewed as vectors may be scaled through multiplication by a real number and two vectors may also be added.

Addition of velocity vectors can be experienced on a bike where the wind velocity and our own velocity relative to the ground add together to form the total velocity relative to the surrounding atmosphere, which is reflected in the air resistance we feel. To compute the total flight time across the Atlantic, the airplane pilot adds the velocity vector of the airplane versus the atmosphere and the velocity of the jet-stream together to obtain the velocity of the airplane vs the ground. We will return below to applications of analytic geometry to mechanics, including these examples.

20.7 Ordered pairs as points or vectors/arrows

We have seen that we may interpret an ordered pair of real numbers (a_1, a_2) as a *point* a in $\mathbb{R}^2$ with coordinates a_1 and a_2. We may write $a = (a_1, a_2)$ for short, and say that a_1 is the first coordinate of the point a and a_2 the second coordinate of a.

We shall also interpret an ordered pair $(a_1, a_2) \in \mathbb{R}^2$ in a alternative way, namely as an *arrow* with tail at the origin and the head at the point $a = (a_1, a_2)$, see Fig. 20.6. With the arrow interpretation of (a_1, a_2), we refer to (a_1, a_2) as a *vector*. Again, we agree to write $a = (a_1, a_2)$, and we say that a_1 and a_2 are the *components* of the arrow/vector $a = (a_1, a_2)$. We say that a_1 is the *first component*, occurring in the first place and a_2 the *second component* occurring in the second place.

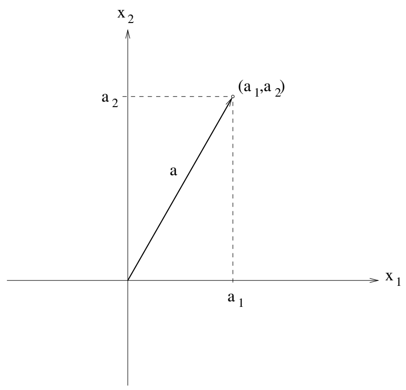


FIGURE 20.6. A vector with tail at the origin and the head at the point $a = (a_1, a_2)$

We thus may interpret an ordered pair (a_1, a_2) in $\mathbb{R}^2$ in two ways: as a point with coordinates (a_1, a_2), or as an arrow/vector with components (a_1, a_2) starting at the origin and ending at the point (a_1, a_2). Evidently, there is a very strong connection between the point and arrow interpretations, since the head of the arrow is located at the point (and assuming that the arrow tail is at the origin). In applications, *positions* will be connected to the point interpretation and *velocities* and *forces* will be connected to the arrow/vector interpretation. We will below generalize the arrow/vector interpretation to include arrows with tails also at other points than the origin. The context will indicate which interpretation is most appropriate for a given situation. Often the interpretation of $a = (a_1, a_2)$ as a point or as an arrow, changes without notice. So we have to be flexible and use whatever interpretation is most convenient or appropriate. We will need even more fantasy when we go into applications to mechanics below.

Sometimes vectors like $a = (a_1, a_2)$ are marked by boldface or an arrow, like $\mathbf{a}$ or $\vec{a}$ or $\underline{a}$, or double script or some other notation. We prefer not to use this more elaborate notation, which makes the writing simpler, but requires fantasy from the user to make the proper interpretation of for example the letter a as a scalar number or vector $a = (a_1, a_2)$ or something else.

20.8 Vector addition

We now proceed to define addition of vectors in $\mathbb{R}^2$, and multiplication of vectors in $\mathbb{R}^2$ by real numbers. In this context, we interpret $\mathbb{R}^2$ as a set of vectors represented by arrows with tail at the origin.

Given two vectors $a = (a_1, a_2)$ and $b = (b_1, b_2)$ in $\mathbb{R}^2$, we use $a + b$ to denote the vector (a_1+b_1, a_2+b_2) in $\mathbb{R}^2$ obtained by adding the components separately. We call $a+b$ the *sum* of a and b obtained through *vector addition*. Thus if $a = (a_1, a_2)$ and $b = (b_1, b_2)$ are given vectors in $\mathbb{R}^2$, then

$$(a_1, a_2) + (b_1, b_2) = (a_1 + b_1, a_2 + b_2), \tag{20.1}$$

which says that vector addition is carried out by adding components separately. We note that $a+b = b+a$ since $a_1+b_1 = b_1+a_1$ and $a_2+b_2 = b_2+a_2$. We say that $0 = (0, 0)$ is the *zero vector* since $a+0 = 0+a = a$ for any vector a. Note the difference between the *vector* zero and its two zero components, which are usually scalars.

EXAMPLE 20.1. We have $(2, 5)+(7, 1) = (9, 6)$ and $(2.1, 5.3)+(7.6, 1.9) = (9.7, 7.2)$.

20.9 Representing vector addition via the Parallelogram Law

We may represent vector addition geometrically using the *Parallelogram Law* as follows. The vector $a+b$ corresponds to the arrow along the diagonal in the parallelogram with two sides formed by the arrows a and b displayed in Fig. 20.7. This follows by noting that the coordinates of the head of $a+b$ is obtained by adding the coordinates of the points a and b separately. This is illustrated in Fig. 20.7.

This definition of vector addition implies that we may reach the point $(a_1 + b_1, a_2 + b_2)$ by walking along arrows in two different ways. First, we simply follow the arrow $(a_1 + b_1, a_2 + b_2)$ to its head, corresponding to walking along the diagonal of the parallelogram formed by a and b. Secondly, we could follow the arrow a from the origin to its head at the point (a_1, a_2) and then continue to the head of the arrow $\bar{b}$ parallel to b and of equal length as b with tail at (a_1, a_2). Alternative, we may follow the arrow b from the origin to its head at the point (b_1, b_2) and then continue to the head of the arrow $\bar{a}$ parallel to a and of equal length as a with tail at (b_1, b_2). The three different routes to the point $(a_1 + b_1, a_2 + b_2)$ are displayed in Fig. 20.7.

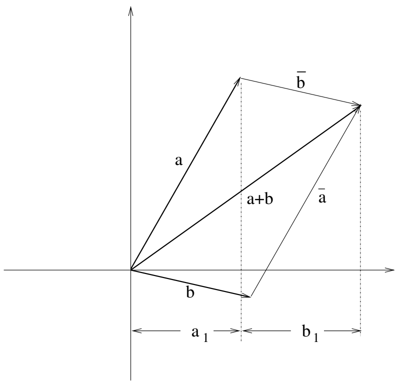


FIGURE 20.7. Vector addition using the Parallelogram Law

We sum up in the following theorem:

Theorem 20.1 *Adding two vectors $a = (a_1, a_2)$ and $b = (b_1, b_2)$ in $\mathbb{R}^2$ to get the sum $a + b = (a_1 + b_1, a_2 + b_2)$ corresponds to adding the arrows a and b using the Parallelogram Law.*

In particular, we can write a vector as the sum of its components in the coordinate directions as follows, see Fig. 20.8.

$$(a_1, a_2) = (a_1, 0) + (0, a_2). \tag{20.2}$$

20.10 Multiplication of a vector by a real number

Given a real number λ and a vector $a = (a_1, a_2) \in \mathbb{R}^2$, we define a new vector $\lambda a \in \mathbb{R}^2$ by

$$\lambda a = \lambda(a_1, a_2) = (\lambda a_1, \lambda a_2). \tag{20.3}$$

For example, $3\,(1.1, 2.3) = (3.3, 6.9)$. We say that λa is obtained by *multiplying* the vector $a = (a_1, a_2)$ by the real number λ and call this operation *multiplication of a vector by a scalar*. Below we will meet other types of multiplication connected with *scalar product of vectors* and *vector product of vectors*, both being different from multiplication of a vector by a scalar.

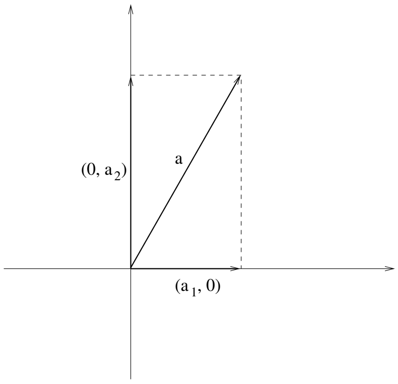


FIGURE 20.8. A vector a represented as the sum of two vectors parallel with the coordinate axes.

We define $-a = (-1)a = (-a_1, -a_2)$ and $a - b = a + (-b)$. We note that $a - a = a + (-a) = (a_1 - a_1, a_2 - a_2) = (0, 0) = 0$. We give an example in Fig. 20.9.

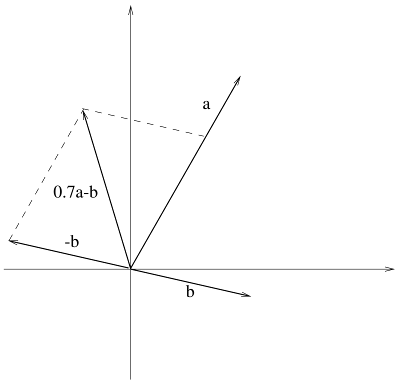


FIGURE 20.9. The sum $0.7a - b$ of the multiples $0.7a$ and $(-1)b$ of a and b.

20.11 The norm of a vector

We define the *Euclidean norm* $|a|$ of a vector $a = (a_1, a_2) \in \mathbb{R}^2$ as

$$|a| = (a_1^2 + a_2^2)^{1/2}. \tag{20.4}$$

By Pythagoras theorem and Fig. 20.10, the Euclidean norm $|a|$ of the vector $a = (a_1, a_2)$ is equal to the length of the hypothenuse of the right angled triangle with sides a_1 and a_2. In other words, the Euclidean norm of the vector $a = (a_1, a_2)$ is equal to the distance from the origin to the point $a = (a_1, a_2)$, or simply the length of the arrow (a_1, a_2). We have $|\lambda a| = |\lambda||a|$ if $\lambda \in \mathbb{R}$ and $a \in \mathbb{R}^2$; multiplying a vector by the real number λ changes the norm of the vector by the factor $|\lambda|$. The zero vector $(0, 0)$ has Euclidean norm 0 and if a vector has Euclidean norm 0 then it must be the zero vector.

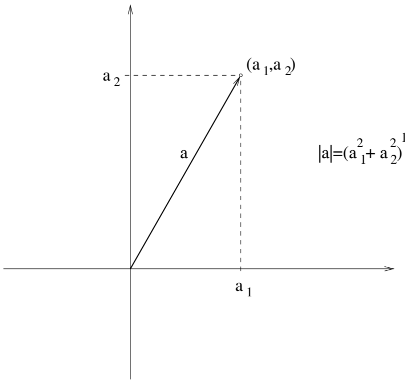


FIGURE 20.10. The norm $|a|$ of a vector $a = (a_1, a_2)$ is $|a| = (a_1^2 + a_2^2)^{1/2}$.

The Euclidean norm of a vector measures the “length” or “size” of the vector. There are many possible ways to measure the “size” of a vector corresponding to using different norms. We will meet several alternative norms of a vector $a = (a_1, a_2)$ below, such as $|a_1| + |a_2|$ or $\max(|a_1|, |a_2|)$. We used $|a_1| + |a_2|$ in the definition of Lipschitz continuity of $f : \mathbb{R}^2 \to \mathbb{R}$ above.

EXAMPLE 20.2. If $a = (3, 4)$ then $|a| = \sqrt{9 + 16} = 5$, and $|2a| = 10$.

20.12 Polar representation of a vector

The points $a = (a_1, a_2)$ in $\mathbb{R}^2$ with $|a| = 1$, corresponding to the vectors a of Euclidean norm equal to 1, form a circle with radius equal to 1 centered at the origin which we call the *unit circle*, see Fig. 20.11.

Each point a on the unit circle can be written $a = (\cos(\theta), \sin(\theta))$ for some angle θ, which we refer to as the *angle of direction* or *direction* of the vector a. This follows from the definition of $\cos(\theta)$ and $\sin(\theta)$ in Chapter Pythagoras and Euclid, see Fig. 20.11

Any vector $a = (a_1, a_2) \neq (0, 0)$ can be expressed as

$$a = |a|\hat{a} = r(\cos(\theta), \sin(\theta)), \tag{20.5}$$

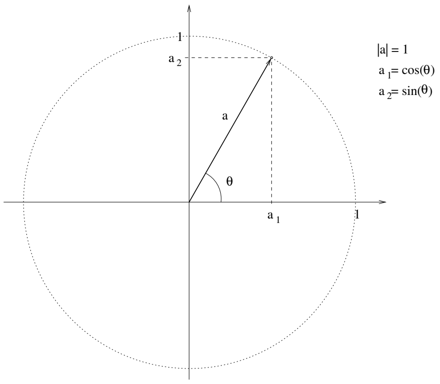


FIGURE 20.11. Vectors of length one are given by $(\cos(\theta), \sin(\theta))$

where $r = |a|$ is the norm of a, $\hat{a} = (a_1/|a|, a_2/|a|)$ is a vector of length one, and θ is the angle of direction of $\hat{a}$, see Fig. 20.12. We call (20.5) the *polar representation* of a. We call θ the direction of a and r the length of a, see Fig. 20.12.

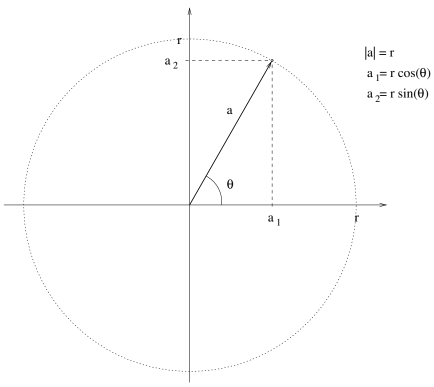


FIGURE 20.12. Vectors of length r are given by $a = r(\cos(\theta), \sin(\theta)) = (r\cos(\theta), r\sin(\theta))$ where $r = |a|$.

We see that if $b = \lambda a$, where $\lambda > 0$ and $a \neq 0$, then b has the same direction as a. If $\lambda < 0$ then b has the opposite direction. In both cases, the norms change with the factor $|\lambda|$; we have $|b| = |\lambda||a|$.

If $b = \lambda a$, where $\lambda \neq 0$ and $a \neq 0$, then we say that the vector b is *parallel to* a. Two parallel vectors have the same or opposite directions.

Example 20.3. We have

$$(1,1) = \sqrt{2}(\cos(45^\circ), \sin(45^\circ)) \quad \text{and} \quad (-1,1) = \sqrt{2}(\cos(135^\circ), \sin(135^\circ)).$$

20.13 Standard basis vectors in $\mathbb{R}^2$

We refer to the vectors $e_1 = (1,0)$ and $e_2 = (0,1)$ as the *standard basis vectors* in $\mathbb{R}^2$. A vector $a = (a_1, a_2)$ can be expressed in term of the basis vectors e_1 and e_2 as

$$a = a_1 e_1 + a_2 e_2,$$

since

$$a_1 e_1 + a_2 e_2 = a_1(1,0) + a_2(0,1) = (a_1, 0) + (0, a_2) = (a_1, a_2) = a.$$

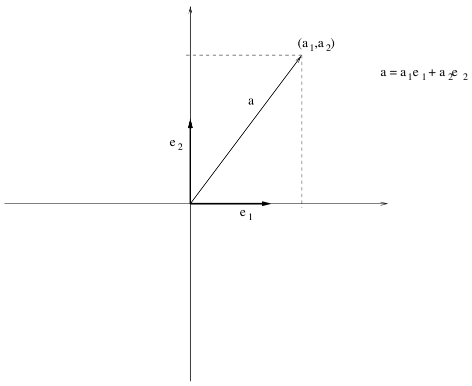


FIGURE 20.13. The standard basis vectors e_1 and e_2 and a linear combination $a = (a_1, a_2) = a_1 e_1 + a_2 e_2$ of e_1 and e_2.

We say that $a_1 e_1 + a_2 e_2$ is a *linear combination* of e_1 and e_2 with *coefficients* a_1 and a_2. Any vector $a = (a_1, a_2)$ in $\mathbb{R}^2$ can thus be expressed as a linear combination of the basis vectors e_1 and e_2 with the coordinates a_1 and a_2 as coefficients, see Fig. 20.13.

EXAMPLE 20.4. We have $(3,7) = 3\,(1,0) + 7\,(0,1) = 3e_1 + 7e_2$.

20.14 Scalar product

While adding vectors to each other and scaling a vector by a real number multiplication have natural interpretations, we shall now introduce a (first) *product of two vectors* that is less motivated at first sight.

Given two vectors $a = (a_1, a_2)$ and $b = (b_1, b_2)$ in $\mathbb{R}^2$, we define their *scalar product* $a \cdot b$ by

$$a \cdot b = a_1 b_1 + a_2 b_2. \tag{20.6}$$

We note, as the terminology suggests, that the scalar product $a \cdot b$ of two vectors a and b in $\mathbb{R}^2$ is a *scalar*, that is a number in $\mathbb{R}$, while the factors a

and b are *vectors* in $\mathbb{R}^2$. Note also that forming the scalar product of two vectors involves not only multiplication, but also a summation!

We note the following connection between the scalar product and the norm:

$$|a| = (a \cdot a)^{\frac{1}{2}}. \tag{20.7}$$

Below we shall define another type of product of vectors where also the product is a vector. We shall thus consider two different types of products of two vectors, which we will refer to as the *scalar product* and the *vector product* respectively. At first when limiting our study to vectors in $\mathbb{R}^2$, we may also view the vector product to be a single real number. However, the vector product in $\mathbb{R}^3$ is indeed a vector in $\mathbb{R}^3$. (Of course, there is also the (trivial) "componentwise" vector product like *MATLAB*© 's $a. * b = (a_1b_1, a_2b_2)$.)

We may view the scalar product as a function $f : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ where $f(a,b) = a \cdot b$. To each pair of vectors $a \in \mathbb{R}^2$ and $b \in \mathbb{R}^2$, we associate the number $f(a,b) = a \cdot b \in \mathbb{R}$. Similarly we may view summation of two vectors as a function $f : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2$. Here, $\mathbb{R}^2 \times \mathbb{R}^2$ denotes the set of all ordered pairs (a,b) of vectors $a = (a_1, a_2)$ and $b = (b_1, b_2)$ in $\mathbb{R}^2$ of course.

EXAMPLE 20.5. We have $(3,7)\cdot(5,2) = 15+14 = 29$, and $(3,7)\cdot(3,7) = 9 + 49 = 58$ so that $|(3,7)| = \sqrt{58}$.

20.15 Properties of the scalar product

The scalar product $a \cdot b$ in $\mathbb{R}^2$ is *linear* in each of the arguments a and b, that is

$$\begin{aligned} a \cdot (b+c) &= a \cdot b + a \cdot c, \\ (a+b) \cdot c &= a \cdot c + b \cdot c, \\ (\lambda a) \cdot b = \lambda\, a \cdot b, &\quad a \cdot (\lambda b) = \lambda\, a \cdot b, \end{aligned}$$

for all $a, b \in \mathbb{R}^2$ and $\lambda \in \mathbb{R}$. This follows directly from the definition (20.6). For example, we have

$$\begin{aligned} a \cdot (b+c) &= a_1(b_1 + c_1) + a_2(b_2 + c_2) \\ &= a_1b_1 + a_2b_2 + a_1c_1 + a_2c_2 = a \cdot b + a \cdot c. \end{aligned}$$

Using the notation $f(a,b) = a \cdot b$, the linearity properties may be written as

$$\begin{aligned} f(a, b+c) = f(a,b) + f(a,c), &\qquad f(a+b, c) = f(a,c) + f(b,c), \\ f(\lambda a, b) = \lambda f(a,b) &\qquad f(a, \lambda b) = \lambda f(a,b). \end{aligned}$$

We also say that the scalar product $a \cdot b = f(a, b)$ is a *bilinear form* on $\mathbb{R}^2 \times \mathbb{R}^2$, that is a function from $\mathbb{R}^2 \times \mathbb{R}^2$ to $\mathbb{R}$, since $a \cdot b = f(a, b)$ is a real number for each pair of vectors a and b in $\mathbb{R}^2$ and $a \cdot b = f(a, b)$ is linear both in the variable (or argument) a and the variable b. Furthermore, the scalar product $a \cdot b = f(a, b)$ is *symmetric* in the sense that

$$a \cdot b = b \cdot a \quad \text{or} \quad f(a, b) = f(b, a),$$

and *positive definite*, that is

$$a \cdot a = |a|^2 > 0 \quad \text{for } a \neq 0 = (0, 0).$$

We may summarize by saying that the scalar product $a \cdot b = f(a, b)$ is a *bilinear symmetric positive definite form on* $\mathbb{R}^2 \times \mathbb{R}^2$.

We notice that for the basis vectors $e_1 = (1, 0)$ and $e_2 = (0, 1)$, we have

$$e_1 \cdot e_2 = 0, \quad e_1 \cdot e_1 = 1, \quad e_2 \cdot e_2 = 1.$$

Using these relations, we can compute the scalar product of two arbitrary vectors $a = (a_1, a_2)$ and $b = (b_1, b_2)$ in $\mathbb{R}^2$ using the linearity as follows:

$$\begin{aligned} a \cdot b &= (a_1 e_1 + a_2 e_2) \cdot (b_1 e_1 + b_2 e_2) \\ &= a_1 b_1\, e_1 \cdot e_1 + a_1 b_2\, e_1 \cdot e_2 + a_2 b_1\, e_2 \cdot e_1 + a_2 b_2\, e_2 \cdot e_2 = a_1 b_1 + a_2 b_2. \end{aligned}$$

We may thus define the scalar product by its action on the basis vectors and then extend it to arbitrary vectors using the linearity in each variable.

20.16 Geometric interpretation of the scalar product

We shall now prove that the scalar product $a \cdot b$ of two vectors a and b in $\mathbb{R}^2$ can be expressed as

$$a \cdot b = |a||b| \cos(\theta), \tag{20.8}$$

where θ is the angle between the vectors a and b, see Fig. 20.14. This formula has a geometric interpretation. Assuming that $|\theta| \leq 90°$ so that $\cos(\theta)$ is positive, consider the right-angled triangle OAC shown in Fig. 20.14. The length of the side OC is $|a| \cos(\theta)$ and thus $a \cdot b$ is equal to the product of the lengths of sides OC and OB. We will refer to OC as the *projection* of OA onto OB, considered as vectors, and thus we may say that $a \cdot b$ is equal to the product of the length of the projection of OA onto OB and the length of OB. Because of the symmetry, we may also relate $a \cdot b$ to the projection of OB onto OA, and conclude that $a \cdot b$ is also equal to the product of the length of the projection of OB onto OA and the length of OA.

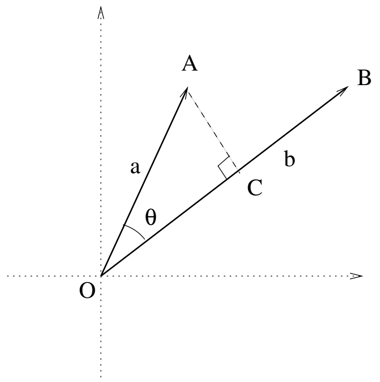


FIGURE 20.14. $a \cdot b = |a|\,|b|\cos(\theta)$.

To prove (20.8), we write using the polar representation

$$a = (a_1, a_2) = |a|(\cos(\alpha), \sin(\alpha)), \qquad b = (b_1, b_2) = |b|(\cos(\beta), \sin(\beta)),$$

where α is the angle of the direction of a and β is the angle of direction of b. Using a basic trigonometric formula from Chapter Pythagoras and Euclid, we see that

$$\begin{aligned} a \cdot b = a_1 b_1 + a_2 b_2 = |a||b|(\cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta)) \\ = |a||b|\cos(\alpha - \beta) = |a||b|\cos(\theta), \end{aligned}$$

where $\theta = \alpha - \beta$ is the angle between a and b. Note that since $\cos(\theta) = \cos(-\theta)$, we may compute the angle between a and b as $\alpha - \beta$ or $\beta - \alpha$.

20.17 Orthogonality and scalar product

We say that two non-zero vectors a and b in $\mathbb{R}^2$ are *geometrically orthogonal*, which we write as $a \perp b$, if the angle between the vectors is 90° or 270°, see Fig. 20.15. The basis vectors e_1 and e_2 are examples of geometrically orthogonal vectors, see Fig. 20.13.

Let a and b be two non-zero vectors making an angle θ. From (20.8), we have $a \cdot b = |a||b|\cos(\theta)$ and thus $a \cdot b = 0$ if and only if $\cos(\theta) = 0$, that is, if and only if $\theta = 90°$ or $\theta = 270°$. We have now proved the following basic result, which we state as a theorem.

Theorem 20.2 *Two non-zero vectors a and b are geometrically orthogonal if and only if $a \cdot b = 0$.*

This result fits our experience in the chapter Pythagoras and Euclid, where we saw that the angle OAB formed by two line segments extending from the origin O out to the points $A = (a_1, a_2)$ and $B = (b_1, b_2)$

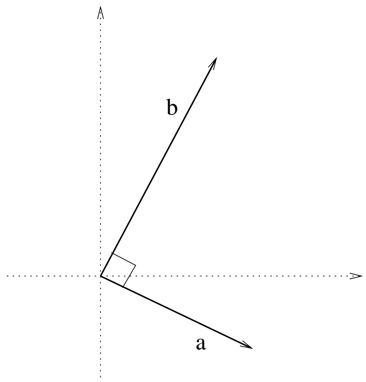


FIGURE 20.15. Orthogonal vectors a and b.

respectively is a right angle if and only if

$$a_1b_1 + a_2b_2 = 0.$$

Summing up, we have translated the *geometric* condition of two vectors $a = (a_1, a_2)$ and $b = (b_1, b_2)$ being geometrically orthogonal to the *algebraic* condition $a \cdot b = a_1b_1 + a_2b_2 = 0$.

Below, in a more general context we will turn this around and *define* two vectors a and b to be *orthogonal* if $a \cdot b = 0$, where $a \cdot b$ is the scalar product of a and b. We have just seen that this algebraic definition of orthogonality may be viewed as an extension of our intuitive idea of geometric orthogonality in $\mathbb{R}^2$. This follows the basic principle of analytic geometry of expressing geometrical relations in algebraic terms.

20.18 Projection of a vector onto a vector

The concept of *projection* is basic in linear algebra. We will now meet this concept for the first time and will use it in many different contexts below.

Let $a = (a_1, a_2)$ and $b = (b_1, b_2)$ be two non-zero vectors and consider the following fundamental problem: Find vectors c and d such that c is parallel to a, d is orthogonal to a, and $c + d = b$, see Fig. 20.16. We refer to $b = c + d$ as an *orthogonal decomposition* of b. We refer to the vector c as *the projection of* b *in the direction of* a, or the *projection of* b *onto* a, and we use the notation $P_a(b) = c$. We can then express the decomposition of b as $b = P_a(b) + (b - P_a(b))$, with $c = P_a(b)$ and $d = b - P_a(b)$. The following properties of the decomposition are immediate:

$$P_a(b) = \lambda a \quad \text{for some } \lambda \in \mathbb{R},$$
$$(b - P_a(b)) \cdot a = 0.$$

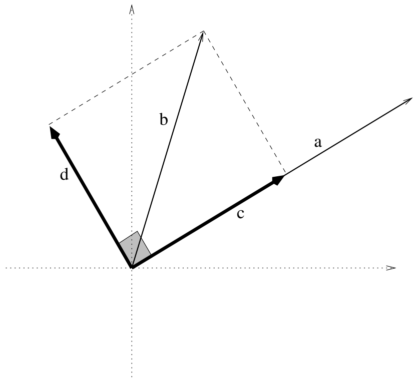


FIGURE 20.16. Orthogonal decomposition of b

Inserting the first equation into the second, we get the equation $(b-\lambda a)\cdot a = 0$ in λ, which we solve to get

$$\lambda = \frac{b \cdot a}{a \cdot a} = \frac{b \cdot a}{|a|^2},$$

and conclude that the projection $P_a(b)$ of b onto a is given by

$$P_a(b) = \frac{b \cdot a}{|a|^2}\, a. \tag{20.9}$$

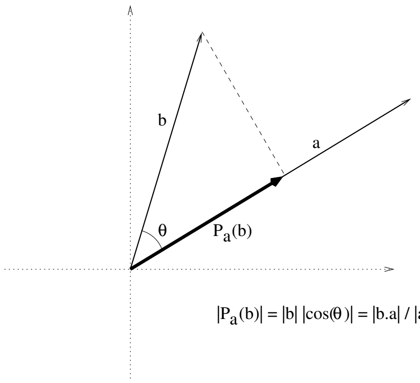


FIGURE 20.17. The projection $P_a(b)$ of b onto a.

We compute the length of $P_a(b)$ as

$$|P_a(b)| = \frac{|a \cdot b|}{|a|^2}|a| = \frac{|a|\,|b|\,|\cos(\theta)|}{|a|} = |b||\cos(\theta)|, \tag{20.10}$$

where θ is the angle between a and b, and we use (20.8). We note that

$$|a \cdot b| = |a|\,|Pb|, \tag{20.11}$$

which conforms with our experience with the scalar product in Section 20.16, see also Fig. 20.15.

We can view the projection $P_a(b)$ of the vector b onto the vector a as a transformation of $\mathbb{R}^2$ into $\mathbb{R}^2$: given the vector $b \in \mathbb{R}^2$, we define the vector $P_a(b) \in \mathbb{R}^2$ by the formula

$$P_a(b) = \frac{b \cdot a}{|a|^2}\, a. \tag{20.12}$$

We write for short $Pb = P_a(b)$, suppressing the dependence on a and the parenthesis, and note that the mapping $P : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $x \to Px$ is linear. We have

$$P(x + y) = Px + Py, \quad P(\lambda x) = \lambda Px, \tag{20.13}$$

for all x and y in $\mathbb{R}^2$ and $\lambda \in \mathbb{R}$ (where we changed name of the independent variable from b to x or y), see Fig. 20.18.

We note that $P(Px) = Px$ for all $x \in \mathbb{R}^2$. This could also be expressed as $P^2 = P$, which is a characteristic property of a projection. Projecting a second time doesn't change anything!

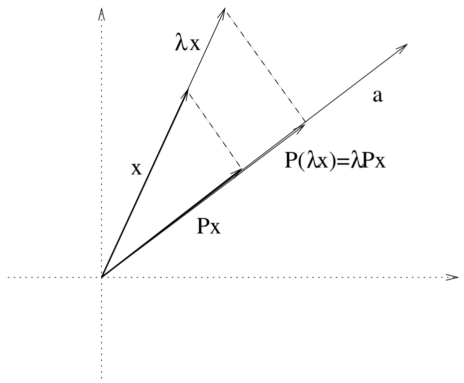


FIGURE 20.18. $P(\lambda x) = \lambda Px$.

We sum up:

Theorem 20.3 *The projection $x \to Px = P_a(x)$ onto a given nonzero vector $a \in \mathbb{R}^2$ is a linear mapping $P : \mathbb{R}^2 \to \mathbb{R}^2$ with the property that $PP = P$.*

EXAMPLE 20.6. If $a = (1, 3)$ and $b = (5, 2)$, then $P_a(b) = \frac{(1,3)\cdot(5,2)}{1+3^2}(1, 3) = (1.1, 3.3)$.

20.19 Rotation by 90°

We saw above that to find the orthogonal decomposition $b = c + d$ with c parallel to a given vector a, it suffices to find c because $d = b - c$. Alternatively, we could seek to first compute d from the requirement that it should be orthogonal to a. We are thus led to the problem of finding a direction orthogonal to a given direction, that is the problem of rotating a given vector by 90°, which we now address.

Given a vector $a = (a_1, a_2)$ in $\mathbb{R}^2$, a quick computation shows that the vector $(-a_2, a_1)$ has the desired property, because computing its scalar product with $a = (a_1, a_2)$ gives

$$(-a_2, a_1) \cdot (a_1, a_2) = (-a_2)a_1 + a_1 a_2 = 0,$$

and thus $(-a_2, a_1)$ is orthogonal to (a_1, a_2). Further, it follows directly that the vector $(-a_2, a_1)$ has the same length as a.

Assuming that $a = |a|(\cos(\alpha), \sin(\alpha))$ and using the facts that $-\sin(\alpha) = \cos(\alpha + 90°)$ and $\cos(\alpha) = \sin(\alpha + 90°)$, we see that the vector $(-a_2, a_1) = |a|(\cos(\alpha + 90°), \sin(\alpha + 90°))$ is obtained by rotating the vector (a_1, a_2) counter-clockwise 90°, see Fig. 20.19. Similarly, the vector $(a_2, -a_1) = -(-a_2, a_1)$ is obtained by clockwise rotation of (a_1, a_2) by 90°.

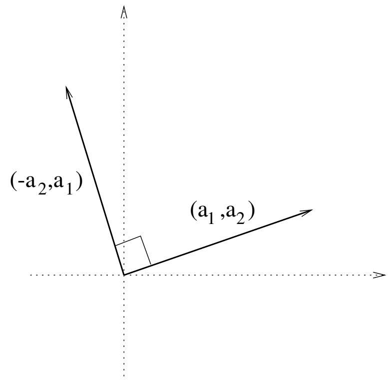


FIGURE 20.19. Counter-clockwise rotation of $a = (a_1, a_2)$ by 90°.

We may view the counter clockwise rotation of a vector by 90° as a *transformation* of vectors: given a vector $a = (a_1, a_2)$, we obtain another vector $a^\perp = f(a)$ through the formula

$$a^\perp = f(a) = (-a_2, a_1),$$

where we denoted the image of the vector a by both $a^\perp$ and $f(a)$. The transformation $a \to a^\perp = f(a)$ defines a linear function $f : \mathbb{R}^2 \to \mathbb{R}^2$ since

$$f(a+b) = (-(a_2+b_2), a_1+b_1) = (-a_2, a_1) + (-b_2, b_1) = f(a) + f(b),$$
$$f(\lambda a) = (-\lambda a_2, \lambda a_1) = \lambda(-a_2, a_1) = \lambda f(a).$$

To specify the action of $a \to a^\perp = f(a)$ on an arbitrary vector a, it suffices to specify the action on the basis vectors e_1 and e_2:

$$e_1^\perp = f(e_1) = (0,1) = e_2, \quad e_2^\perp = f(e_2) = (-1,0) = -e_1,$$

since by linearity, we may compute

$$\begin{aligned} a^\perp = f(a) &= f(a_1e_1 + a_2e_2) = a_1 f(e_1) + a_2 f(e_2) \\ &= a_1(0,1) + a_2(-1,0) = (-a_2, a_1). \end{aligned}$$

EXAMPLE 20.7. Rotating the vector $(1,2)$ the angle $90°$ counter-clockwise, we get the vector $(-2,1)$.

20.20 Rotation by an arbitrary angle θ

We now generalize to counter-clockwise rotation by an arbitrary angle θ. Let $a = |a|(\cos(\alpha), \sin(\alpha))$ in $\mathbb{R}^2$ be a given vector. We seek a vector $R_\theta(a)$ in $\mathbb{R}^2$ of equal length obtained by rotating a the angle θ counter-clockwise. By the definition of the vector $R_\theta(a)$ as the vector $a = |a|(\cos(\alpha), \sin(\alpha))$ rotated by θ, we have

$$R_\theta(a) = |a|(\cos(\alpha+\theta), \sin(\alpha+\theta)).$$

Using the standard trigonometric formulas from Chapter Pythagoras and Euclid,

$$\cos(\alpha+\theta) = \cos(\alpha)\cos(\theta) - \sin(\alpha)\sin(\theta),$$
$$\sin(\alpha+\theta) = \sin(\alpha)\cos(\theta) + \cos(\alpha)\sin(\theta),$$

we can write the formula for the rotated vector $R_\theta(a)$ as

$$R_\theta(a) = (a_1\cos(\theta) - a_2\sin(\theta), a_1\sin(\theta) + a_2\cos(\theta)). \tag{20.14}$$

We may view the counter-clockwise rotation of a vector by the angle θ as a *transformation* of vectors: given a vector $a = (a_1, a_2)$, we obtain another vector $R_\theta(a)$ by rotation by θ according to the above formula. Of course, we may view this transformation as a function $R_\theta : \mathbb{R}^2 \to \mathbb{R}^2$. It is easy to verify that this function is linear. To specify the action of R_θ on an arbitrary vector a, it suffices to specify the action on the basis vectors e_1 and e_2,

$$R_\theta(e_1) = (\cos(\theta), \sin(\theta)), \quad R_\theta(e_2) = (-\sin(\theta), \cos(\theta)).$$

The formula (20.14) may then be obtained using linearity,

$$\begin{aligned} R_\theta(a) &= R_\theta(a_1e_1 + a_2e_2) = a_1R_\theta(e_1) + a_2R_\theta(e_2) \\ &= a_1(\cos(\theta), \sin(\theta)) + a_2(-\sin(\theta), \cos(\theta)) \\ &= (a_1\cos(\theta) - a_2\sin(\theta), a_1\sin(\theta) + a_2\cos(\theta)). \end{aligned}$$

EXAMPLE 20.8. Rotating the vector $(1, 2)$ the angle $30°$, we obtain the vector $(\cos(30°) - 2\sin(30°), \sin(30°) + 2\cos(30°)) = (\frac{\sqrt{3}}{2} - 1, \frac{1}{2} + \sqrt{3})$.

20.21 Rotation by θ again!

We present yet another way to arrive at (20.14) based on the idea that the transformation $R_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ of counter-clockwise rotation by θ is defined by the following properties,

$$\text{(i)} \quad |R_\theta(a)| = |a|, \quad \text{and} \quad \text{(ii)} \quad R_\theta(a) \cdot a = \cos(\theta)|a|^2. \tag{20.15}$$

Property (i) says that rotation preserves the length and (ii) connects the change of direction to the scalar product. We now seek to determine $R_\theta(a)$ from (i) and (ii). Given $a \in \mathbb{R}^2$, we set $a^\perp = (-a_2, a_1)$ and express $R_\theta(a)$ as $R_\theta(a) = \alpha a + \beta a^\perp$ with appropriate real numbers α and β. Taking the scalar product with a and using $a \cdot a^\perp = 0$, we find from (ii) that $\alpha = \cos(\theta)$. Next, (i) states that $|a|^2 = |R_\theta(a)|^2 = (\alpha^2 + \beta^2)|a|^2$, and we conclude that $\beta = \pm\sin(\theta)$ and thus finally $\beta = \sin(\theta)$ using the counter-clockwise orientation. We conclude that

$$R_\theta(a) = \cos(\theta)a + \sin(\theta)a^\perp = (a_1\cos(\theta) - a_2\sin(\theta), a_1\sin(\theta) + a_2\cos(\theta)),$$

and we have recovered (20.14).

20.22 Rotating a coordinate system

Suppose we rotate the standard basis vectors $e_1 = (1, 0)$ and $e_2 = (0, 1)$ counter-clockwise the angle θ to get the new vectors $\hat{e}_1 = \cos(\theta)e_1 + \sin(\theta)e_2$ and $\hat{e}_2 = -\sin(\theta)e_1 + \cos(\theta)e_2$. We may use $\hat{e}_1$ and $\hat{e}_2$ as an alternative coordinate system, and we may seek the connection between the coordinates of a given vector (or point) in the old and new coordinate system. Letting

(a_1, a_2) be the coordinates in the standard basis e_1 and e_2, and $(\hat{a}_1, \hat{a}_2)$ the coordinates in the new basis $\hat{e}_1$ and $\hat{e}_2$, we have

$$\begin{aligned} a_1 e_1 + a_2 e_2 = \hat{a}_1 \hat{e}_1 + \hat{a}_2 \hat{e}_2 & \\ &= \hat{a}_1(\cos(\theta)e_1 + \sin(\theta)e_2) + \hat{a}_2(-\sin(\theta)e_1 + \cos(\theta)e_2) \\ &= (\hat{a}_1\cos(\theta) - \hat{a}_2\sin(\theta))e_1 + (\hat{a}_1\sin(\theta) + \hat{a}_2\cos(\theta))e_2, \end{aligned}$$

so the uniqueness of coordinates with respect to e_1 and e_2 implies

$$\begin{aligned} a_1 &= \cos(\theta)\hat{a}_1 - \sin(\theta)\hat{a}_2, \\ a_2 &= \sin(\theta)\hat{a}_1 + \cos(\theta)\hat{a}_2. \end{aligned} \tag{20.16}$$

Since e_1 and e_2 are obtained by rotating $\hat{e}_1$ and $\hat{e}_2$ clockwise by the angle θ,

$$\begin{aligned} \hat{a}_1 &= \cos(\theta)a_1 + \sin(\theta)a_2, \\ \hat{a}_2 &= -\sin(\theta)a_1 + \cos(\theta)a_2. \end{aligned} \tag{20.17}$$

The connection between the coordinates with respect to the two coordinate systems is thus given by (20.16) and (20.17).

Example 20.9. Rotating 45° counter-clockwise gives the following relation between new and old coordinates

$$\hat{a}_1 = \frac{1}{\sqrt{2}}(a_1 + a_2), \quad \hat{a}_2 = \frac{1}{\sqrt{2}}(-a_1 + a_2).$$

20.23 Vector product in $\mathbb{R}^2$

We now define the *vector product* $a \times b$ of two vectors $a = (a_1, a_2)$ and $b = (b_1, b_2)$ in $\mathbb{R}^2$ by the formula

$$a \times b = a_1 b_2 - a_2 b_1. \tag{20.18}$$

The vector product $a \times b$ is also referred to as the *cross product* because of the notation used (don't mix up with the "$\times$" in the "product set" $\mathbb{R}^2 \times \mathbb{R}^2$ which has a different meaning). The vector product or cross product may be viewed as a function $\mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$. This function is bilinear as is easy to verify, and *anti-symmetric*, that is

$$a \times b = -b \times a, \tag{20.19}$$

which is a surprising property for a product.

Since the vector product is bilinear, we can specify the action of the vector product on two arbitrary vectors a and b by specifying the action on the basis vectors,

$$e_1 \times e_1 = 0,\ e_2 \times e_2 = 0,\ e_1 \times e_2 = 1,\ e_2 \times e_1 = -1. \tag{20.20}$$

Using these relations,

$$a \times b = (a_1e_1 + a_2e_2) \times (b_1e_1 + b_2e_2) = a_1b_2e_1 \times e_2 + a_2b_1e_2 \times e_1 = a_1b_2 - a_2b_1e_2.$$

We next show that the properties of bilinearity and anti-symmetry in fact determine the vector product in $\mathbb{R}^2$ up to a constant. First note that anti-symmetry and bilinearity imply

$$\begin{aligned} e_1 \times e_1 + e_1 \times e_2 = e_1 \times (e_1 + e_2) = -(e_1 + e_2) \times e_1 \\ = -e_1 \times e_1 - e_2 \times e_1. \end{aligned}$$

Since $e_1 \times e_2 = -e_2 \times e_1$, we have $e_1 \times e_1 = 0$. Similarly, we see that $e_2 \times e_2 = 0$. We conclude that the action of the vector product on the basis vectors is indeed specified according to (20.20) up to a constant.

We next observe that

$$a \times b = (-a_2, a_1) \cdot (b_1, b_2) = a_1b_2 - a_2b_1,$$

which gives a connection between the vector product $a \times b$ and the scalar product $a^\perp \cdot b$ with the $90°$ counter-clockwise rotated vector $a^\perp = (-a_2, a_1)$. We conclude that the vector product $a \times b$ of two nonzero vectors a and b is zero if and only if a and b are parallel. We state this basic result as a theorem.

Theorem 20.4 *Two nonzero vectors a and b are parallel if and only if $a \times b = 0$.*

We can thus check if two non-zero vectors a and b are parallel by checking if $a \times b = 0$. This is another example of translating a geometric condition (two vectors a and b being parallel) into an algebraic condition ($a \times b = 0$).

We now squeeze more information from the relation $a \times b = a^\perp \cdot b$ assuming that the angle between a and b is θ and thus the angle between $a^\perp$ and b is $\theta + 90°$:

$$\begin{aligned} |a \times b| = |a^\perp \cdot b| &= |a^\perp||b|\,|\cos(\theta + \frac{\pi}{2})| \\ &= |a|\,|b|\,|\sin(\theta)|, \end{aligned}$$

where we use $|a^\perp| = |a|$ and $|\cos(\theta \pm \pi/2)| = |\sin(\theta)|$. Therefore,

$$|a \times b| = |a||b||\sin(\theta)|, \tag{20.21}$$

where $\theta = \alpha - \beta$ is the angle between a and b, see Fig. 20.20

We can make the formula (20.21) more precise by removing the absolute values around $a \times b$ and the sine factor if we adopt a suitable sign convention. This leads to the following more developed version of (20.21), which we state as a theorem, see Fig. 20.21.

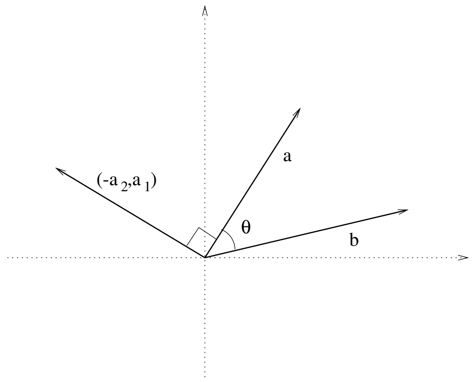


FIGURE 20.20. Why $|a \times b| = |a|\,|b|\,|\sin(\theta)|$.

Theorem 20.5 *For two non-zero vectors a and b,*

$$a \times b = |a||b|\sin(\theta), \tag{20.22}$$

where θ is the angle between a and b counted positive counter-clockwise and negative clockwise starting from a.

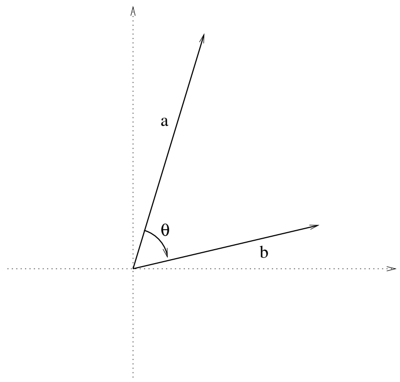


FIGURE 20.21. $a \times b = |a|\,|b|\sin(\theta)$ is negative here because the angle θ is negative.

20.24 The area of a triangle with a corner at the origin

Consider a triangle OAB with corners at the origin O and the points $A = (a_1, a_2)$ and $B = (b_1, b_2)$ formed by the vectors $a = (a_1, a_2)$ and $b = (b_1, b_2)$, see Fig. 20.22. We say that the triangle OAB is *spanned* by the vectors a and b. We are familiar with the formula that states that the area of this triangle can be computed as the base $|a|$ times the height $|b||\sin(\theta)|$ times the factor $\frac{1}{2}$, where θ is the angle between a and b, see Fig. 20.22. Recalling (20.21), we conclude

Theorem 20.6

$$Area(OAB) = \frac{1}{2}|a|\,|b|\,|\sin(\theta)| = \frac{1}{2}\,|a \times b|.$$

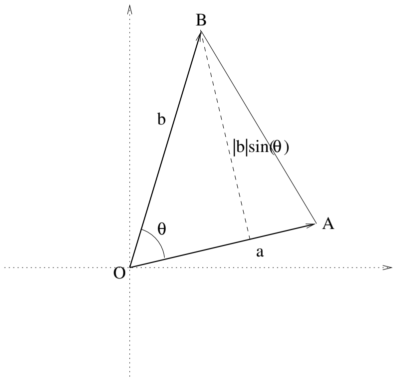


FIGURE 20.22. The vectors a and b span a triangle with area $\frac{1}{2}|a \times b|$.

The area of the triangle OAB can be computed using the vector product in $\mathbb{R}^2$.

20.25 The area of a general triangle

Consider a triangle CAB with corners at the points $C = (c_1, c_2)$, $A = (a_1, a_2)$ and $B = (b_1, b_2)$. We consider the problem of computing the area of the triangle CAB. We solved this problem above in the case $C = O$ where O is the origin. We may reduce the present case to that case by changing coordinate system as follows. Consider a new coordinate system with origin at $C = (c_1, c_2)$ and with a $\hat{x}_1$-axis parallel to the x_1-axis and a $\hat{x}_2$-axis parallel to the x_2-axis, see Fig. 20.23.

Letting $(\hat{a}_1, \hat{a}_2)$ denote the coordinates with respect to the new coordinate system, the new are related to the old coordinates by

$$\hat{a}_1 = a_1 - c_1, \quad \hat{a}_2 = a_2 - c_2.$$

The coordinates of the points A, B and C in the new coordinate system are thus $(a_1 - c_1, a_2 - c_2) = a - c$, $(b_1 - c_1, b_2 - c_2) = b - c$ and $(0, 0)$. Using

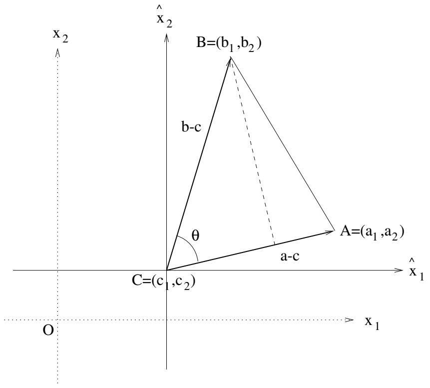


FIGURE 20.23. Vectors $a-c$ and $b-c$ span triangle with area $\frac{1}{2}|(a-c)\times(b-c)|$.

the result from the previous section, we find the area of the triangle CAB by the formula

$$\text{Area}(CAB) = \frac{1}{2}|(a-c)\times(b-c)|. \tag{20.23}$$

EXAMPLE 20.10. The area of the triangle with coordinates at $A = (2,3)$, $B = (-2,2)$ and $C = (1,1)$, is given by $\text{Area}(CAB) = \frac{1}{2}|(1,2)\times(-3,1)| = \frac{7}{2}$.

20.26 The area of a parallelogram spanned by two vectors

The area of the parallelogram spanned by a and b, as shown in Fig. 20.24, is equal to $|a\times b|$ since the area of the parallelogram is twice the area of the triangle spanned by a and b. Denoting the area of the parallelogram spanned by the vectors a and b by $V(a,b)$, we thus have the formula

$$V(a,b) = |a\times b|. \tag{20.24}$$

This is a fundamental formula which has important generalizations to $\mathbb{R}^3$ and $\mathbb{R}^n$.

20.27 Straight lines in $\mathbb{R}^2$

The points $x = (x_1, x_2)$ in the plane $\mathbb{R}^2$ satisfying a relation of the form

$$n_1x_1 + n_2x_2 = n\cdot x = 0, \tag{20.25}$$

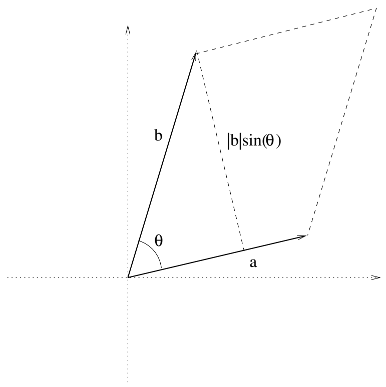


FIGURE 20.24. The vectors a and b span a rectangle with area $|a \times b| = |a|\,|b|\,\sin(\theta)|$.

where $n = (n_1, n_2) \in \mathbb{R}^2$ is a given non-zero vector, form a straight line through the origin that is orthogonal to (n_1, n_2), see Fig. 20.25. We say that (n_1, n_2) is a *normal* to the line. We can represent the points $x \in \mathbb{R}^2$ on the line in the form

$$x = sn^{\perp}, \quad s \in \mathbb{R},$$

where $n^{\perp} = (-n_2, n_1)$ is orthogonal to n, see Fig. 20.25. We state this insight as a theorem because of its importance.

Theorem 20.7 *A line in $\mathbb{R}^2$ passing through the origin with normal $n \in \mathbb{R}^2$, may be expressed as either the points $x \in \mathbb{R}^2$ satisfying $n \cdot x = 0$, or the set of points of the form $x = sn^{\perp}$ with $n^{\perp} \in \mathbb{R}^2$ orthogonal to n and $s \in \mathbb{R}$.*

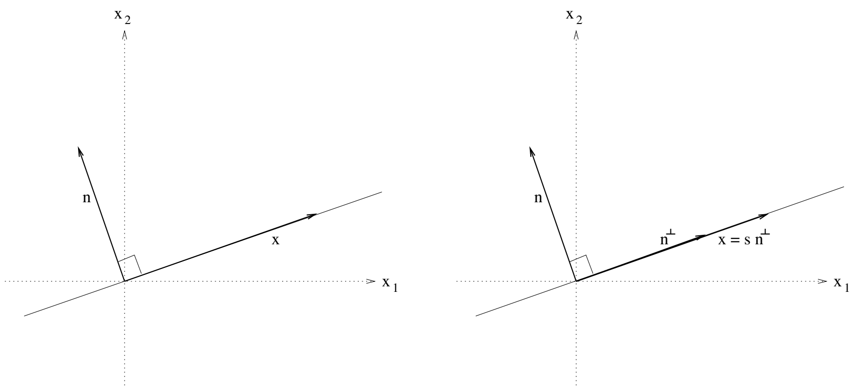


FIGURE 20.25. Vectors $x = sa$ with b orthogonal to a given vector n generate a line through the origin with normal a.

Similarly, the set of points (x_1, x_2) in $\mathbb{R}^2$ such that

$$n_1 x_1 + n_2 x_2 = d, \tag{20.26}$$

where $n = (n_1, n_2) \in \mathbb{R}^2$ is a given non-zero vector and d is a given constant, represents a straight line that does not pass through the origin if $d \neq 0$. We see that n is a normal to the line, since if x and $\hat{x}$ are two points on the line then $(x - \hat{x}) \cdot n = d - d = 0$, see Fig. 20.26. We may define the line as the points $x = (x_1, x_2)$ in $\mathbb{R}^2$, such that the projection $\frac{n \cdot x}{|n|^2} n$ of the vector $x = (x_1, x_2)$ in the direction of n is equal to $\frac{d}{|n|^2} n$. To see this, we use the definition of the projection and the fact $n \cdot x = d$.

The line in Fig. 20.25 can also be represented as the set of points

$$x = \hat{x} + s n^{\perp} \quad s \in \mathbb{R},$$

where $\hat{x}$ is any point on the line (thus satisfying $n \cdot \hat{x} = d$). This is because any point x of the form $x = s n^{\perp} + \hat{x}$ evidently satisfies $n \cdot x = n \cdot \hat{x} = d$. We sum up in the following theorem.

Theorem 20.8 *The set of points $x \in \mathbb{R}^2$ such that $n \cdot x = d$, where $n \in \mathbb{R}^2$ is a given non-zero vector and d is given constant, represents a straight line in $\mathbb{R}^2$. The line can also be expressed in the form $x = \hat{x} + s n^{\perp}$ for $s \in \mathbb{R}$, where $\hat{x} \in \mathbb{R}^2$ is a point on the line.*

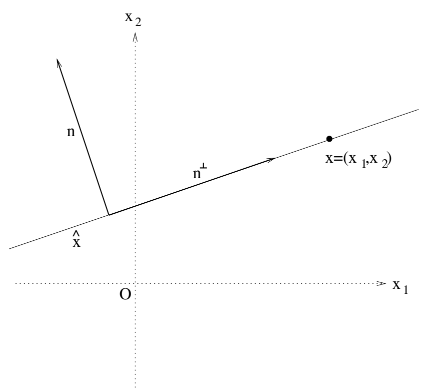


FIGURE 20.26. The line through the point $\hat{x}$ with normal n generated by directional vector a.

EXAMPLE 20.11. The line $x_1 + 2x_2 = 3$ can alternatively be expressed as the set of points $x = (1, 1) + s(-2, 1)$ with $s \in \mathbb{R}$.

20.28 Projection of a point onto a line

Let $n \cdot x = d$ represent a line in $\mathbb{R}^2$ and let b be a point in $\mathbb{R}^2$ that does not lie on the line. We consider the problem of finding the point Pb on the line which is closest to b, see Fig. 20.27. This is called the *projection* of the point b onto the line. Equivalently, we seek a point Pb on the line such that $b - Pb$ is orthogonal to the line, that is we seek a point Pb such that

$$n \cdot Pb = d \quad (Pb \text{ is a point on the line}),$$
$$b - Pb \text{ is parallel to the normal } n, \quad (b - Pb = \lambda n, \quad \text{for some } \lambda \in \mathbb{R}).$$

We conclude that $Pb = b - \lambda n$ and the equation $n \cdot Pb = d$ thus gives $n \cdot (b - \lambda n) = d$, that is $\lambda = \frac{b \cdot n - d}{|n|^2}$ and so

$$Pb = b - \frac{b \cdot n - d}{|n|^2} n. \tag{20.27}$$

If $d = 0$, that is the line $n \cdot x = d = 0$ passes through the origin, then (see Problem 20.26)

$$Pb = b - \frac{b \cdot n}{|n|^2} n. \tag{20.28}$$

20.29 When are two lines parallel?

Let

$$a_{11}x_1 + a_{12}x_2 = b_1,$$
$$a_{21}x_1 + a_{22}x_2 = b_2,$$

be two straight lines in $\mathbb{R}^2$ with normals (a_{11}, a_{12}) and (a_{21}, a_{22}). How do we know if the lines are parallel? Of course, the lines are parallel if and only if their normals are parallel. From above, we know the normals are parallel if and only if

$$(a_{11}, a_{12}) \times (a_{21}, a_{22}) = a_{11}a_{22} - a_{12}a_{21} = 0,$$

and consequently *non*-parallel (and consequently intersecting at some point) if and only if

$$(a_{11}, a_{12}) \times (a_{21}, a_{22}) = a_{11}a_{22} - a_{12}a_{21} \neq 0, \tag{20.29}$$

EXAMPLE 20.12. The two lines $2x_1 + 3x_2 = 1$ and $3x_1 + 4x_2 = 1$ are non-parallel because $2 \cdot 4 - 3 \cdot 3 = 8 - 9 = -1 \neq 0$.

20.30 A system of two linear equations in two unknowns

If $a_{11}x_1 + a_{12}x_2 = b_1$ and $a_{21}x_1 + a_{22}x_2 = b_2$ are two straight lines in $\mathbb{R}^2$ with normals (a_{11}, a_{12}) and (a_{21}, a_{22}), then their intersection is determined by the *system of linear equations*

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1, \\ a_{21}x_1 + a_{22}x_2 &= b_2, \end{aligned} \tag{20.30}$$

which says that we seek a point $(x_1, x_2) \in \mathbb{R}^2$ that lies on both lines. This is a *system of two linear equations in two unknowns* x_1 and x_2, or a 2×2-system. The numbers a_{ij}, $i, j = 1, 2$ are the *coefficients* of the system and the numbers b_i, $i = 1, 2$, represent the given *right hand side*.

If the normals (a_{11}, a_{12}) and (a_{21}, a_{22}) are not parallel or by (20.29), $a_{11}a_{22} - a_{12}a_{21} \neq 0$, then the lines must intersect and thus the system (20.30) should have a unique solution (x_1, x_2). To determine x_1, we multiply the first equation by a_{22} to get

$$a_{11}a_{22}x_1 + a_{12}a_{22}x_2 = b_1a_{22}.$$

We then multiply the second equation by a_{12}, to get

$$a_{21}a_{12}x_1 + a_{22}a_{12}x_2 = b_2a_{12}.$$

Subtracting the two equations the x_2-terms cancel and we get the following equation containing only the unknown x_1,

$$a_{11}a_{22}x_1 - a_{21}a_{12}x_1 = b_1a_{22} - b_2a_{12}.$$

Solving for x_1, we get

$$x_1 = (a_{22}b_1 - a_{12}b_2)(a_{11}a_{22} - a_{12}a_{21})^{-1}.$$

Similarly to determine x_2, we multiply the first equation by a_{21} and subtract the second equation multiplied by a_{11}, which eliminates a_1. Altogether, we obtain the solution formula

$$x_1 = (a_{22}b_1 - a_{12}b_2)(a_{11}a_{22} - a_{12}a_{21})^{-1}, \tag{20.31a}$$
$$x_2 = (a_{11}b_2 - a_{21}b_1)(a_{11}a_{22} - a_{12}a_{21})^{-1}. \tag{20.31b}$$

This formula gives the unique solution of (20.30) under the condition $a_{11}a_{22} - a_{12}a_{21} \neq 0$.

We can derive the solution formula (20.31) in a different way, still assuming that $a_{11}a_{22} - a_{12}a_{21} \neq 0$. We define $a_1 = (a_{11}, a_{21})$ and $a_2 = (a_{12}, a_{22})$, noting carefully that here a_1 and a_2 denote *vectors* and that $a_1 \times a_2 =$

$a_{11}a_{22} - a_{12}a_{21} \neq 0$, and rewrite the two equations of the system (20.30) in vector form as

$$x_1 a_1 + x_2 a_2 = b. \tag{20.32}$$

Taking the vector product of this equation with a_2 and a_1 and using $a_2 \times a_2 = a_1 \times a_1 = 0$,

$$x_1 a_1 \times a_2 = b \times a_2, \quad x_2 a_2 \times a_1 = b \times a_1.$$

Since $a_1 \times a_2 \neq 0$,

$$x_1 = \frac{b \times a_2}{a_1 \times a_2}, \quad x_2 = \frac{b \times a_1}{a_2 \times a_1} = -\frac{b \times a_1}{a_1 \times a_2}, \tag{20.33}$$

which agrees with the formula (20.31) derived above.

We conclude this section by discussing the case when $a_1 \times a_2 = a_{11}a_{22} - a_{12}a_{21} = 0$, that is the case when a_1 and a_2 are parallel or equivalently the two lines are parallel. In this case, $a_2 = \lambda a_1$ for some $\lambda \in \mathbb{R}$ and the system (20.30) has a solution if and only if $b_2 = \lambda b_1$, since then the second equation results from multiplying the first by λ. In this case there are infinitely many solutions since the two lines coincide. In particular if we choose $b_1 = b_2 = 0$, then the solutions consist of all (x_1, x_2) such that $a_{11}x_1 + a_{12}x_2 = 0$, which defines a straight line through the origin. On the other hand if $b_2 \neq \lambda b_1$, then the two equations represent two different parallel lines that do not intersect and there is no solution to the system (20.30).

We summarize our experience from this section on systems of 2 linear equations in 2 unknowns as follows:

Theorem 20.9 *The system of linear equations $x_1 a_1 + x_2 a_2 = b$, where a_1, a_2 and b are given vectors in $\mathbb{R}^2$, has a unique solution (x_1, x_2) given by (20.33) if $a_1 \times a_2 \neq 0$. In the case $a_1 \times a_2 = 0$, the system has no solution or infinitely many solutions, depending on b.*

Below we shall generalize this result to systems of n linear equations in n unknowns, which represents one of the most basic results of linear algebra.

EXAMPLE 20.13. The solution to the system

$$\begin{aligned} x_1 + 2x_2 &= 3, \\ 4x_1 + 5x_2 &= 6, \end{aligned}$$

is given by

$$x_1 = \frac{(3,6) \times (2,5)}{(1,4) \times (2,5)} = \frac{3}{-3} = -1, \quad x_2 = -\frac{(3,6) \times (1,4)}{(1,4) \times (2,5)} = -\frac{6}{-3} = 2.$$

20.31 The concepts of linear independence and basis

We saw above that the system (20.30) can be written in vector form as

$$x_1a_1 + x_2a_2 = b,$$

where $b = (b_1, b_2)$, $a_1 = (a_{11}, a_{21})$ and $a_2 = (a_{12}, a_{22})$ are all vectors in $\mathbb{R}^2$, and x_1 and x_2 real numbers. We say that

$$x_1a_1 + x_2a_2,$$

is a *linear combination* of the vectors a_1 and a_2, or a linear combination of the set of vectors $\{a_1, a_2\}$, with the coefficients x_1 and x_2 being real numbers. The system of equations (20.30) expresses the right hand side vector b as a linear combination of the set of vectors $\{a_1, a_2\}$ with the coefficients x_1 and x_2. We refer to x_1 and x_2 as the *coordinates* of b with respect to the set of vectors $\{a_1, a_2\}$, which we may write as an ordered pair (x_1, x_2).

The solution formula (20.33) thus states that if $a_1 \times a_2 \neq 0$, then an arbitrary vector b in $\mathbb{R}^2$ can be expressed as a linear combination of the set of vectors $\{a_1, a_2\}$ with the coefficients x_1 and x_2 being uniquely determined. This means that if $a_1 \times a_2 \neq 0$, then the the set of vectors $\{a_1, a_2\}$ may serve as a *basis* for $\mathbb{R}^2$, in the sense that each vector b in $\mathbb{R}^2$ may be uniquely expressed as a linear combination $b = x_1a_1 + x_2a_2$ of the set of vectors $\{a_1, a_2\}$. We say that the ordered pair (x_1, x_2) are the *coordinates* of b with respect to the basis $\{a_1, a_2\}$. The system of equations $b = x_1a_1 + x_2a_2$ thus give the coupling between the coordinates (b_1, b_2) of the vector b in the standard basis, and the coordinates (x_1, x_2) with respect to the basis $\{a_1, a_2\}$. In particular, if $b = 0$ then $x_1 = 0$ and $x_2 = 0$.

Conversely if $a_1 \times a_2 = 0$, that is a_1 and a_2 are parallel, then any nonzero vector b orthogonal to a_1 is also orthogonal to a_2 and b cannot be expressed as $b = x_1a_1 + x_2a_2$. Thus, if $a_1 \times a_2 = 0$ then $\{a_1, a_2\}$ cannot serve as a basis. We have now proved the following basic theorem:

Theorem 20.10 *A set $\{a_1, a_2\}$ of two non-zero vectors a_1 and a_2 may serve as a basis for $\mathbb{R}^2$ if and only if if $a_1 \times a_2 \neq 0$. The coordinates (b_1, b_2) of a vector b in the standard basis and the coordinates (x_1, x_2) of b with respect to a basis $\{a_1, a_2\}$ are related by the system of linear equations $b = x_1a_1 + x_2a_2$.*

EXAMPLE 20.14. The two vectors $a_1 = (1, 2)$ and $a_2 = (2, 1)$ (expressed in the standard basis) form a basis for $\mathbb{R}^2$ since $a_1 \times a_2 = 1 - 4 = -3$. Let $b = (5, 4)$ in the standard basis. To express b in the basis $\{a_1, a_2\}$, we seek real numbers x_1 and x_2 such that $b = x_1a_1 + x_2a_2$, and using the solution formula (20.33) we find that $x_1 = 1$ and $x_2 = 2$. The coordinates of b with respect to the basis $\{a_1, a_2\}$ are thus $(1, 2)$, while the coordinates of b with respect to the standard basis are $(5, 4)$.

We next introduce the concept of *linear independence*, which will play an important role below. We say that a set $\{a_1, a_2\}$ of two non-zero vectors a_1 and a_2 two non-zero vectors a_1 and a_2 in $\mathbb{R}^2$ is *linearly independent* if the system of equations

$$x_1 a_1 + x_2 a_2 = 0$$

has the unique solution $x_1 = x_2 = 0$. We just saw that if $a_1 \times a_2 \neq 0$, then a_1 and a_2 are linearly independent (because $b = (0, 0)$ implies $x_1 = x_2 = 0$). Conversely if $a_1 \times a_2 = 0$, then a_1 and a_2 are parallel so that $a_1 = \lambda a_2$ for some $\lambda \neq 0$, and then there are many possible choices of x_1 and x_2, not both equal to zero, such that $x_1 a_1 + x_2 a_2 = 0$, for example $x_1 = -1$ and $x_2 = \lambda$. We have thus proved:

Theorem 20.11 *The set $\{a_1, a_2\}$ of non-zero vectors a_1 and a_2 is linearly independent if and only if $a_1 \times a_2 \neq 0$.*

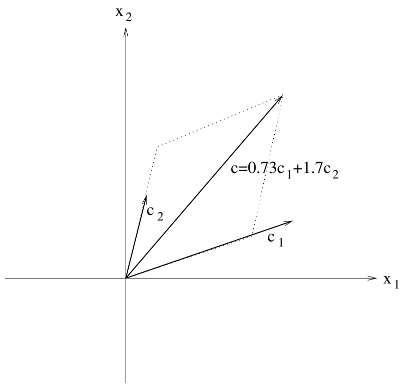


FIGURE 20.27. Linear combination c of two linearly independent vectors c_1 and c_2

20.32 The connection to calculus in one variable

We have discussed Calculus of real-valued functions $y = f(x)$ of one real variable $x \in \mathbb{R}$, and we have used a coordinate system in $\mathbb{R}^2$ to plot graphs of functions $y = f(x)$ with x and y representing the two coordinate axis. Alternatively, we may specify the graph as the set of points $(x_1, x_2) \in \mathbb{R}^2$, consisting of pairs (x_1, x_2) of real numbers x_1 and x_2, such that $x_2 = f(x_1)$ or $x_2 - f(x_1) = 0$ with x_1 representing x and x_2 representing y. We refer to the ordered pair $(x_1, x_2) \in \mathbb{R}^2$ as a vector $x = (x_1, x_2)$ with components x_1 and x_2.

We have also discussed properties of linear functions $f(x) = ax + b$, where a and b are real constants, the graphs of which are straight lines $x_2 = ax_1 + b$ in $\mathbb{R}^2$. More generally, a straight line in $\mathbb{R}^2$ is the set of points $(x_1, x_2) \in \mathbb{R}^2$ such that $x_1 a_1 + x_2 a_2 = b$, where the a_1, a_2 and b are real constants, with $a_1 \neq 0$ and/or $a_2 \neq 0$. We have noticed that (a_1, a_2) may be viewed as a direction in $\mathbb{R}^2$ that is perpendicular or normal to the line $a_1 x_1 + a_2 x_2 = b$, and that $(b/a_1, 0)$ or $(0, b/a_2)$ are the points where the line intersects the x_1-axis and the x_2-axis respectively.

20.33 Linear mapping $f : \mathbb{R}^2 \to \mathbb{R}$

A function $f : \mathbb{R}^2 \to \mathbb{R}$ is *linear* if for any $x = (x_1, x_2)$ and $y = (y_1, y_2)$ in $\mathbb{R}^2$ and any λ in $\mathbb{R}$,

$$f(x + y) = f(x) + f(y) \quad \text{and} \quad f(\lambda x) = \lambda f(x). \tag{20.34}$$

Setting $c_1 = f(e_1) \in \mathbb{R}$ and $c_2 = f(e_2) \in \mathbb{R}$, where $e_1 = (1, 0)$ and $e_2 = (0, 1)$ are the standard basis vectors in $\mathbb{R}^2$, we can represent $f : \mathbb{R}^2 \to \mathbb{R}$ as follows:

$$f(x) = x_1 c_1 + x_2 c_2 = c_1 x_1 + c_2 x_2,$$

where $x = (x_1, x_2) \in \mathbb{R}^2$. We also refer to a linear function as a linear *mapping*.

EXAMPLE 20.15. The function $f(x_1, x_2) = x_1 + 3x_2$ defines a linear mapping $f : \mathbb{R}^2 \to \mathbb{R}$.

20.34 Linear mapping $f : \mathbb{R}^2 \to \mathbb{R}^2$

A function $f : \mathbb{R}^2 \to \mathbb{R}^2$ taking values $f(x) = (f_1(x), f_2(x)) \in \mathbb{R}^2$ is *linear* if the component functions $f_1 : \mathbb{R}^2 \to \mathbb{R}$ and $f_2 : \mathbb{R}^2 \to \mathbb{R}$ are linear. Setting $a_{11} = f_1(e_1)$, $a_{12} = f_1(e_2)$, $a_{21} = f_2(e_1)$, $a_{22} = f_2(e_2)$, we can represent $f : \mathbb{R}^2 \to \mathbb{R}^2$ as $f(x) = (f_1(x), f_2(x))$, where

$$f_1(x) = a_{11} x_1 + a_{12} x_2, \tag{20.35a}$$

$$f_2(x) = a_{21} x_1 + a_{22} x_2, \tag{20.35b}$$

and the a_{ij} are real numbers.

A linear mapping $f : \mathbb{R}^2 \to \mathbb{R}^2$ maps (parallel) lines onto (parallel) lines since for $x = \hat{x} + sb$ and f linear, we have $f(x) = f(\hat{x} + sb) = f(\hat{x}) + sf(b)$, see Fig. 20.28.

EXAMPLE 20.16. The function $f(x_1, x_2) = (x_1 + 3x_2, 2x_1 - x_3)$ defines a linear mapping $\mathbb{R}^2 \to \mathbb{R}^2$.

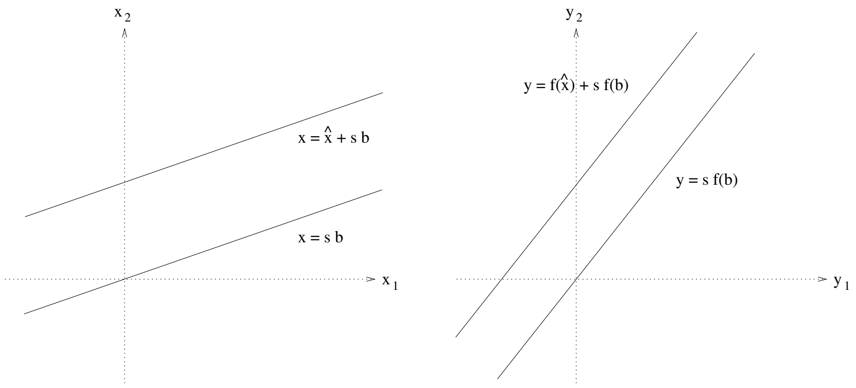


FIGURE 20.28. A linear mapping $f : \mathbb{R}^2 \to \mathbb{R}^2$ maps (parallel) lines to (parallel) lines, and consequently parallelograms to parallelograms.

20.35 Linear mapping $f : \mathbb{R}^2 \to \mathbb{R}^2$ and linear system of equations

Let a linear mapping $f : \mathbb{R}^2 \to \mathbb{R}^2$ and a vector $b \in \mathbb{R}^2$ be given. We consider the problem of finding $x \in \mathbb{R}^2$ such that

$$f(x) = b.$$

Assuming $f(x)$ is represented by (20.35), we seek $x \in \mathbb{R}^2$ satisfying the 2×2 linear system of equations

$$a_{11}x_1 + a_{12}x_2 = b_1, \tag{20.36a}$$

$$a_{21}x_1 + a_{22}x_2 = b_2, \tag{20.36b}$$

where the coefficients a_{ij} and the coordinates b_i of the right hand side are given.

20.36 A first encounter with 2×2 and 2×1 matrices

We write the left hand side of (20.36) as follows:

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{pmatrix}. \tag{20.37}$$

The quadratic array

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

is called a 2×2 *matrix*. We can view this matrix to consist of two rows

$$\begin{pmatrix} a_{11} & a_{12} \end{pmatrix} \qquad \text{and} \qquad \begin{pmatrix} a_{21} & a_{22}, \end{pmatrix}$$

or two columns

$$\begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix}.$$

Each row may be viewed as a 1×2 matrix with 1 horizontal array with 2 elements and each column may be viewed as a 2×1 matrix with 1 vertical array with 2 elements. In particular, the array

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

may be viewed as a 2×1 matrix. We also refer to a 2×1 matrix as a 2-*column vector*, and a 1×2 matrix as a 2-*row vector*. Writing $x = (x_1, x_2)$ we may view x as a 1×2 matrix or 2-row vector. Using matrix notation, it is most natural to view $x = (x_1, x_2)$ as a 2-column vector.

The expression (20.37) defines the *product* of a 2×2 matrix and a 2×1 matrix or a 2-column vector. The product can be interpreted as

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} c_1 \cdot x \\ c_2 \cdot x \end{pmatrix} \tag{20.38}$$

where we interpret $r_1 = (a_{11}, a_{12})$ and $r_2 = (a_{21}, a_{22})$ as the two ordered pairs corresponding to the two rows of the matrix and x is the ordered pair (x_1, x_2). The matrix-vector product is given by

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \tag{20.39}$$

i.e. by taking the scalar product of the ordered pairs r_1 and r_2 corresponding to the 2-row vectors of the matrix with the order pair corresponding to the 2-column vector $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$.

Writing

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \quad \text{and } x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \text{and } b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}, \tag{20.40}$$

we can phrase the system of equations (20.36) in condensed form as the following *matrix equation:*

$$Ax = b, \qquad \text{or} \qquad \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

We have now got a first glimpse of matrices including the basic operation of multiplication of a 2×2-matrix with a 2×1 matrix or 2-column vector.

Below we will generalize to a calculus for matrices including addition of matrices, multiplication of matrices with a real number, and multiplication of matrices. We will also discover a form of matrix division referred to as inversion of matrices allowing us to express the solution of the system $Ax = b$ as $x = A^{-1}b$, under the condition that the columns (or equivalently, the rows) of A are linearly independent.

20.37 First applications of matrix notation

To show the usefulness of the matrix notation just introduced, we rewrite some of the linear systems of equations and transformations which we have met above.

Rotation by θ

The mapping $R_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ corresponding to rotation of a vector by an angle θ is given by (20.14), that is

$$R_\theta(x) = (x_1 \cos(\theta) - x_2 \sin(\theta), x_1 \sin(\theta) + x_2 \cos(\theta)). \tag{20.41}$$

Using matrix notation, we can write $R_\theta(x)$ as follows

$$R_\theta(x) = Ax = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

where A thus is the 2×2 matrix

$$A = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}. \tag{20.42}$$

Projection onto a vector a

The projection $P_a(x) = \frac{x \cdot a}{|a|^2} a$ given by (20.9) of a vector $x \in \mathbb{R}^2$ onto a given vector $a \in \mathbb{R}^2$ can be expressed in matrix form as follows:

$$P_a(x) = Ax = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

where A is the 2×2 matrix

$$A = \begin{pmatrix} \frac{a_1^2}{|a|^2} & \frac{a_1 a_2}{|a|^2} \\ \frac{a_1 a_2}{|a|^2} & \frac{a_2^2}{|a|^2} \end{pmatrix}. \tag{20.43}$$

Change of basis

The linear system (20.17) describing a change of basis can be written in matrix form as

$$\begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

or in condensed from as $\hat{x} = Ax$, where A is the matrix

$$A = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}$$

and x and $\hat{x}$ are 2-column vectors.

20.38 Addition of matrices

Let A be a given 2×2 matrix with elements a_{ij}, $i, j = 1, 2$, that is

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

We write $A = (a_{ij})$. Let $B = (b_{ij})$ be another 2×2 matrix. We define the sum $C = A + B$ to be the matrix $C = (c_{ij})$ with elements $c_{ij} = a_{ij} + b_{ij}$ for $i, j = 1, 2$. In other words, we add two matrices element by element:

$$A + B = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{pmatrix} = C.$$

20.39 Multiplication of a matrix by a real number

Given a 2×2 matrix A with elements a_{ij}, $i, j = 1, 2$, and a real number λ, we define the matrix $C = \lambda A$ as the matrix with elements $c_{ij} = \lambda a_{ij}$. In other words, all elements a_{ij} are multiplied by λ:

$$\lambda A = \lambda \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = \begin{pmatrix} \lambda a_{11} & \lambda a_{12} \\ \lambda a_{21} & \lambda a_{22} \end{pmatrix} = C.$$

20.40 Multiplication of two matrices

Given two 2×2 matrices $A = (a_{ij})$ and $B = (b_{ij})$ with elements, we define the product $C = AB$ as the matrix with elements c_{ij} given by

$$c_{ij} = \sum_{k=1}^{2} a_{ik} b_{kj}.$$

Writing out the sum, we have

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

In other words, to get the element c_{ij} of the product $C = AB$, we take the scalar product of row i of A with column j of B.

The matrix product is generally non-commutative so that $AB \neq BA$ most of the time.

We say that in the product AB the matrix A multiplies the matrix B from the left and that B multiplies the matrix A from the right. Non-commutativity of matrix multiplication means that multiplication from right or left may give different results.

Example 20.17. We have

$$\begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 3 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 5 \\ 2 & 4 \end{pmatrix}, \quad \text{while} \quad \begin{pmatrix} 1 & 3 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 4 & 5 \\ 2 & 3 \end{pmatrix}.$$

Example 20.18. We compute $BB = B^2$, where B is the projection matrix given by (20.43), that is

$$B = \begin{pmatrix} \frac{a_1^2}{|a|^2} & \frac{a_1 a_2}{|a|^2} \\ \frac{a_1 a_2}{|a|^2} & \frac{a_2^2}{|a|^2} \end{pmatrix} = \frac{1}{|a|^2} \begin{pmatrix} a_1^2 & a_1 a_2 \\ a_1 a_2 & a_2^2 \end{pmatrix}.$$

We have

$$BB = \frac{1}{|a|^4} \begin{pmatrix} a_1^2 & a_1 a_2 \\ a_1 a_2 & a_2^2 \end{pmatrix} \begin{pmatrix} a_1^2 & a_1 a_2 \\ a_1 a_2 & a_2^2 \end{pmatrix}$$

$$= \frac{1}{|a|^4} \begin{pmatrix} a_1^2(a_1^2 + a_2^2) & a_1 a_2(a_1^2 + a_2^2) \\ a_1 a_2(a_1^2 + a_2^2) & a_2^2(a_1^2 + a_2^2) \end{pmatrix} = \frac{1}{|a|^2} \begin{pmatrix} a_1^2 & a_1 a_2 \\ a_1 a_2 & a_2^2 \end{pmatrix} = B,$$

and see as expected that $BB = B$.

Example 20.19. As another application we compute the product of two matrices corresponding to two rotations with angles α and β:

$$A = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} \cos(\beta) & -\sin(\beta) \\ \sin(\beta) & \cos(\beta) \end{pmatrix}. \tag{20.44}$$

We compute

$$AB = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix} \begin{pmatrix} \cos(\beta) & -\sin(\beta) \\ \sin(\beta) & \cos(\beta) \end{pmatrix}$$

$$\begin{pmatrix} \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta) & -\cos(\alpha)\sin(\beta) - \sin(\alpha)\cos(\beta) \\ \cos(\alpha)\sin(\beta) + \sin(\alpha)\cos(\beta) & \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta) \end{pmatrix}$$

$$= \begin{pmatrix} \cos(\alpha+\beta) & -\sin(\alpha+\beta) \\ \sin(\alpha+\beta) & \cos(\alpha+\beta) \end{pmatrix},$$

where again we have used the formulas for $\cos(\alpha+\beta)$ and $\sin(\alpha+\beta)$ from Chapter Pythagoras and Euclid. We conclude as expected that two successive rotations of angles α and β corresponds to a rotation of angle $\alpha+\beta$.

20.41 The transpose of a matrix

Given a 2×2 matrix A with elements a_{ij}, we define the *transpose* of A denoted by $A^\top$ as the matrix $C = A^\top$ with elements $c_{11} = a_{11}$, $c_{12} = a_{21}$, $c_{21} = a_{12}$, $c_{22} = a_{22}$. In other words, the rows of A are the columns of $A^\top$ and vice versa. For example

$$\text{if} \quad A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \quad \text{then} \quad A^\top = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}.$$

Of course $(A^\top)^\top = A$. Transposing twice brings back the original matrix. We can directly check the validity of the following rules for computing with the transpose:

$$(A+B)^\top = A^\top + B^\top, \quad (\lambda A)^\top = \lambda A^\top,$$
$$(AB)^\top = B^\top A^\top.$$

20.42 The transpose of a 2-column vector

The transpose of a 2-column vector is the row vector with the same elements:

$$\text{if} \quad x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \text{then} \quad x^\top = \begin{pmatrix} x_1 & x_2 \end{pmatrix}.$$

We may define the product of a 1×2 matrix (2-row vector) $x^\top$ with a 2×1 matrix (2-column vector) y in the natural way as follows:

$$x^\top y = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = x_1 y_1 + x_2 y_2.$$

In particular, we may write

$$|x|^2 = x \cdot x = x^\top x,$$

where we interpret x as an ordered pair and as a 2-column vector.

20.43 The identity matrix

The 2×2 matrix

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

is called the *identity matrix* and is denoted by I. We have $IA = A$ and $AI = A$ for any 2×2 matrix A.

20.44 The inverse of a matrix

Let A be a 2×2 matrix with elements a_{ij} with $a_{11}a_{22} - a_{12}a_{21} \neq 0$. We define the *inverse* matrix A^{-1} by

$$A^{-1} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}. \tag{20.45}$$

We check by direct computation that $A^{-1}A = I$ and that $AA^{-1} = I$, which is the property we ask an "inverse" to satisfy. We get the first column of A^{-1} by using the solution formula (20.31) with $b = (1, 0)$ and the second column choosing $b = (0, 1)$.

The solution to the system of equations $Ax = b$ can be written as $x = A^{-1}b$, which we obtain by multiplying $Ax = b$ from the left by A^{-1}.

We can directly check the validity of the following rules for computing with the inverse:

$$(\lambda A)^{-1} = \lambda A^{-1}$$
$$(AB)^{-1} = B^{-1}A^{-1}.$$

20.45 Rotation in matrix form again!

We have seen that a rotation of a vector x by an angle θ into a vector y can be expressed as $y = R_\theta x$ with R_θ being the rotation matrix:

$$R_\theta = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \tag{20.46}$$

We have also seen that two successive rotations by angles α and β can be written as

$$y = R_\beta R_\alpha x, \tag{20.47}$$

and we have also shown that $R_\beta R_\alpha = R_{\alpha+\beta}$. This states the obvious fact that two successive rotations α and β can be performed as one rotation with angle $\alpha + \beta$.

We now compute the inverse R_θ^{-1} of a rotation R_θ using (20.45),

$$R_\theta^{-1} = \frac{1}{\cos(\theta)^2+\sin(\theta)^2}\begin{pmatrix}\cos(\theta) & \sin(\theta)\\ -\sin(\theta) & \cos(\theta)\end{pmatrix} = \begin{pmatrix}\cos(-\theta) & -\sin(-\theta)\\ \sin(-\theta) & \cos(-\theta)\end{pmatrix} \tag{20.48}$$

where we use $\cos(\alpha)=\cos(-\alpha)$, $\sin(\alpha)=-\sin(-\alpha)$. We see that $R_\theta^{-1} = R_{-\theta}$, which is one way of expressing the (obvious) fact that the inverse of a rotation by θ is a rotation by $-\theta$.

We observe that $R_\theta^{-1} = R_\theta^\top$ with $R_\theta^\top$ the transpose of R_θ, so that in particular

$$R_\theta R_\theta^\top = I. \tag{20.49}$$

We use this fact to prove that the length of a vector is not changed by rotation. If $y = R_\theta x$, then

$$|y|^2 = y^T y = (R_\theta x)^\top(R_\theta x) = x^\top R_\theta^\top R_\theta x = x^\top x = |x|^2. \tag{20.50}$$

More generally, the scalar product is preserved after the rotation. If $y = R_\theta x$ and $\hat{y} = R_\theta \hat{x}$, then

$$y\cdot\hat{y} = (R_\theta x)^\top(R_\theta \hat{x}) = x^\top R_\theta^\top R_\theta \hat{x} = x\cdot\hat{x}. \tag{20.51}$$

The relation (20.49) says that the matrix R_θ is *orthogonal.* Orthogonal matrices play an important role, and we will return to this topic below.

20.46 A mirror in matrix form

Consider the linear transformation $2P-I$, where $Px = \frac{a\cdot x}{|a|^2}a$ is the projection onto the non-zero vector $a\in\mathbb{R}^2$, that is onto the line $x = sa$ through the origin. In matrix form, this can be expressed as

$$2P - I = \frac{2}{|a|^2}\begin{pmatrix}a_1^2-1 & a_1a_2\\ a_2a_1 & a_2^2-1\end{pmatrix}.$$

After some reflection(!), looking at Fig. 20.29, we understand that the transformation $I + 2(P-I) = 2P-I$ maps a point x into its mirror image in the line through the origin with direction a.

To see if $2P-I$ preserves scalar products, we assume that $y=(2P-I)x$ and $\hat{y} = (2P-I)\hat{x}$ and compute:

$$y\cdot\hat{y} = ((2P-I)x)^\top(2P-I)\hat{x} = x^\top(2P^\top - I)(2P-I)\hat{x} = \tag{20.52}$$

$$x^\top(4P^\top P - 2P^\top I - 2PI + I)\hat{x} = x^\top(4P-4P+I)\hat{x} = x\cdot\hat{x}, \tag{20.53}$$

where we used the fact that $P = P^\top$ and $PP = P$, and we thus find an affirmative answer.

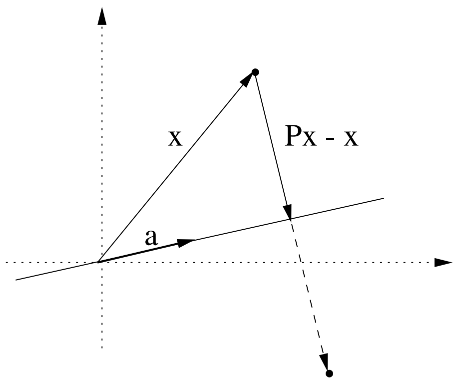


FIGURE 20.29. The mapping $2P - I$ maps points to its mirror point relative to the given line.

20.47 Change of basis again!

Let $\{a_1, a_2\}$ and $\{\hat{a}_1, \hat{a}_2\}$ be two different bases in $\mathbb{R}^2$. We then express any given $b \in \mathbb{R}^2$ as

$$b = x_1 a_1 + x_2 a_2 = \hat{x}_1 \hat{a}_1 + \hat{x}_2 \hat{a}_2, \tag{20.54}$$

with certain coordinates (x_1, x_2) with respect to $\{a_1, a_2\}$ and some other coordinates $(\hat{x}_1, \hat{x}_2)$ with respect $\{\hat{a}_1, \hat{a}_2\}$.

To connect (x_1, x_2) to $(\hat{x}_1, \hat{x}_2)$, we express the basis vectors $\{\hat{a}_1, \hat{a}_2\}$ in terms of the basis $\{a_1, a_2\}$:

$$\begin{aligned} c_{11}a_1 + c_{21}a_2 &= \hat{a}_1, \\ c_{12}a_1 + c_{22}a_2 &= \hat{a}_2, \end{aligned}$$

with certain coefficients c_{ij}. Inserting this into (20.54), we get

$$\hat{x}_1(c_{11}a_1 + c_{21}a_2) + \hat{x}_2(c_{12}a_1 + c_{22}a_2) = b.$$

Reordering terms,

$$(c_{11}\hat{x}_1 + c_{12}\hat{x}_2)a_1 + (c_{21}\hat{x}_1 + c_{22}\hat{x}_2)a_2 = b.$$

We conclude by uniqueness that

$$x_1 = c_{11}\hat{x}_1 + c_{12}\hat{x}_2, \tag{20.55}$$
$$x_2 = c_{21}\hat{x}_1 + c_{22}\hat{x}_2, \tag{20.56}$$

which gives the connection between the coordinates (x_1, x_2) and $(\hat{x}_1, \hat{x}_2)$. Using matrix notation, we can write this relation as $x = C\hat{x}$ with

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}.$$

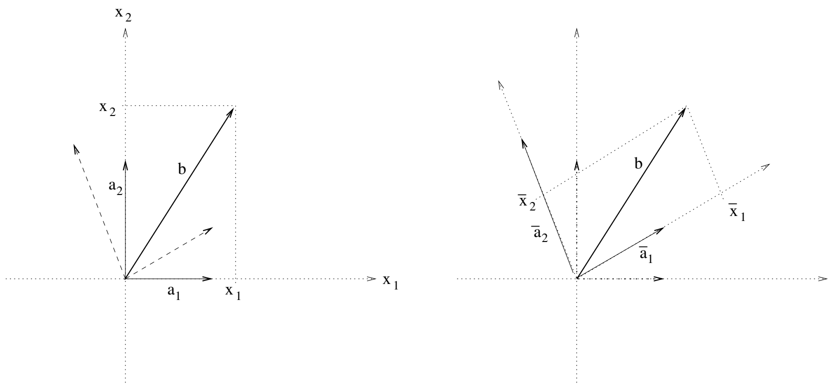


FIGURE 20.30. A vector b may be expressed in terms of the basis $\{a_1, a_2\}$ or the basis $\{\hat{a}_1, \hat{a}_2\}$.

20.48 Queen Christina

Queen Christina of Sweden (1626-1689), daughter of Gustaf Vasa King of Sweden 1611-1632, crowned to Queen at the age 5, officially coronated 1644, abdicated 1652, converted to Catholicism and moved to Rome 1655.

Throughout her life, Christina had a passion for the arts and for learning, and surrounded herself with musicians, writers, artists and also philosophers, theologians, scientists and mathematicians. Christina had an impressive collection of sculpture and paintings, and was highly respected for both her artistic and literary tastes. She also wrote several books, including her *Letters to Descartes* and *Maxims*. Her home, the Palace Farnese, was the active center of cultural and intellectual life in Rome for several decades.

Duc de Guise quoted in Georgina Masson's Queen Christina biography describes Queen Chistina as follows: "She isn't tall, but has a well-filled figure and a large behind, beautiful arms, white hands. One shoulder is higher than another, but she hides this defect so well by her bizarre dress, walk and movements.... The shape of her face is fair but framed by the most extraordinary coiffure. It's a man's wig, very heavy and piled high in front, hanging thick at the sides, and at the back there is some slight resemblance to a woman's coiffure.... She is always very heavily powdered over a lot of face cream".

Chapter 20 Problems

FIGURE 20.31. Queen Christina to Descartes:'If we conceive the world in that vast extension you give it, it is impossible that man conserve himself therein in this honorable rank, on the contrary, he shall consider himself along with the entire earth he inhabits as in but a small, tiny and in no proportion to the enormous size of the rest. He will very likely judge that these stars have inhabitants, or even that the earths surrounding them are all filled with creatures more intelligent and better than he, certainly, he will lose the opinion that this infinite extent of the world is made for him or can serve him in any way".

20.1. Given the vectors a, b och c in $\mathbb{R}^2$ and the scalars $\lambda, \mu \in \mathbb{R}$, prove the following statements

$$a + b = b + a, \quad (a + b) + c = a + (b + c), \quad a + (-a) = 0$$
$$a + 0 = a, \quad 3a = a + a + a, \quad \lambda(\mu a) = (\lambda\mu)a,$$
$$(\lambda + \mu)a = \lambda a + \mu a, \quad \lambda(a + b) = \lambda a + \lambda b, \quad |\lambda a| = |\lambda||a|.$$

Try to give both analytical and geometrical proofs.

20.2. Give a formula for the transformation $f : \mathbb{R}^2 \to \mathbb{R}^2$ corresponding to reflection through the direction of a given vector $a \in \mathbb{R}^2$. Find the corresponding matrix.

20.3. Given $a = (3, 2)$ and $b = (1, 4)$, compute (i) $|a|$, (ii) $|b|$, (iii) $|a + b|$, (iv)) $|a - b|$, (v) $a/|a|$, (vi) $b/|b|$.

20.4. Show that the norm of $a/|a|$ with $a \in \mathbb{R}^2$, $a \neq 0$, is equal to 1.

20.5. Given $a, b \in \mathbb{R}^2$ prove the following inequalities a) $|a + b| \leq |a| + |b|$, b) $a \cdot b \leq |a||b|$.

20.6. Compute $a \cdot b$ with (i) $a = (1, 2), b = (3, 2)$ (ii) $a = (10, 27), b = (14, -5)$

20.7. Given $a, b, c \in \mathbb{R}^2$, determine which of the following statements make sense: (i) $a \cdot b$, (ii) $a \cdot (b \cdot c)$, (iii) $(a \cdot b) + |c|$, (iv) $(a \cdot b) + c$, (v) $|a \cdot b|$.

20.8. What is the angle between $a = (1, 1)$ and $b = (3, 7)$?

20.9. Given $b = (2, 1)$ construct the set of all vectors $a \in \mathbb{R}^2$ such that $a \cdot b = 2$. Give a geometrical interpretation of this result.

20.10. Find the projection of a onto b onto $(1, 2)$ with (i) $a = (1, 2)$, (ii) $a = (-2, 1)$, (iii) $a = (2, 2)$, (iv) $a = (\sqrt{2}, \sqrt{2})$.

20.11. Decompose the vector $b = (3, 5)$ into one component parallel to a and one component orthogonal to a for all vectors a in the previous exercise.

20.12. Let a, b and $c = a - b$ in $\mathbb{R}^2$ be given, and let the angle between a and b be φ. Show that:

$$|c|^2 = |a|^2 + |b|^2 - 2|a||b| \cos \varphi.$$

Give an interpretation of the result.

20.13. Prove the law of cosines for a triangle with sidelengths a, b and c:

$$c^2 = a^2 + b^2 - 2ab \cos(\theta),$$

where θ is the angle between the sides a and b.

20.14. Given the 2 by 2 matrix:

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$$

compute Ax and $A^T x$ for the following choice of $x \in \mathbb{R}^2$: (i) $x^T = (1, 2)$ (ii) $x^T = (1, 1)$

20.15. Given the 2×2-matrices:

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, B = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix},$$

compute (i) AB, (ii) BA, (iii) $A^T B$, (iv) AB^T, (v) $B^T A^T$, (vi) $(AB)^T$, (vii) A^{-1}, (viii) B^{-1}, (ix) $(AB)^{-1}$, (x) $A^{-1}A$.

20.16. Show that $(AB)^T = B^T A^T$ and that $(Ax)^T = x^T A^T$.

20.17. What can be said about A if: a) $A = A^T$ b) $AB = I$?

20.18. Show that the projection:

$$P_a(b) = \frac{b \cdot a}{|a|^2} a$$

can be written in the form Pb, where P is a 2×2 matrix. Show that $PP = P$ and $P = P^T$.

20.19. Compute the mirror image of a point with respect to a straight line in $\mathbb{R}^2$ which does not pass through the origin. Express the mapping in matrix form.

20.20. Express the linear transformation of rotating a vector a certain given angle as a matrix vector product.

20.21. Given $a, b \in \mathbb{R}^2$, show that the "mirror vector" $\bar{b}$ obtained by reflecting b in a can be expressed as:

$$\bar{b} = 2Pb - b$$

where P is a certain projection Show that the scalar product between two vectors is invariant under a reflection, that is

$$c \cdot d = \bar{c} \cdot \bar{d}.$$

20.22. Compute $a \times b$ and $b \times a$ with (i) $a = (1, 2), b = (3, 2)$, (ii) $a = (1, 2), b = (3, 6)$, (iii) $a = (2, -1), b = (2, 4)$.

20.23. Extend the Matlab functions for vectors in $\mathbb{R}^2$ by writing functions for vector product (x = vecProd(a, b)) and rotation (b = vecRotate(a, angle)) of vectors.

20.24. Check the answers to the above problems using Matlab.

20.25. Verify that the projection $Px = P_a(x)$ is linear in x. Is it linear also in a? Illustrate, as in Fig. 20.18, that $P_a(x + y) = P_a(x) + P_a(y)$.

20.26. Prove that the formula (21.29) for the projection of a point onto a line through the origin, coincides with the formula (20.9) for the projection of the vector b on the direction of the line.

20.27. Show that if $\hat{a} = \lambda a$, where a is a nonzero vector $\mathbb{R}^2$ and $\lambda \neq 0$, then for any $b \in \mathbb{R}^2$ we have $P_{\hat{a}}(b) = P_a(b)$, where $P_a(b)$ is the projection of b onto a. Conclude that the projection onto a non-zero vector $a \in \mathbb{R}^2$ only depends on the direction of a and not the norm of a.

-

21
Analytic geometry in $\mathbb{R}^3$

> We must confess that in all humility that, while number is a product of our mind alone, space has a reality beyond the mind whose rules we cannot completely prescribe. (Gauss 1830)

21.1 Introduction

We now extend the discussion of analytic geometry to *Euclidean three dimensional space* or Euclidean 3d space for short. We imagine this space arises when we draw a normal through the origin to a Euclidean two dimensional plane spanned by orthogonal x_1 and x_2 axes, and call the normal the x_3-axis. We then obtain an orthogonal coordinate system consisting of three coordinate x_1, x_2 and x_3 axes that intersect at the origin, with each axis being a copy of $\mathbb{R}$, see Fig. 21.1.

In daily life, we may imagine a room where we live as a portion of $\mathbb{R}^3$, with the horizontal floor being a piece of $\mathbb{R}^2$ with two coordinates (x_1, x_2) and with the vertical direction as the third coordinate x_3. On a larger scale, we may imagine our neighborhood in terms of three orthogonal directions West-East, South-North, and Down-Up, which may be viewed to be a portion of $\mathbb{R}^3$, if we neglect the curvature of the Earth.

The coordinate system can be oriented two ways, right or left. The coordinate system is said to be right-oriented, which is the standard, if turning a standard screw *into* the direction of the positive x_3-axis will turn the x_1-axis the shortest route to the x_2-axis, see Fig. 21.2. Alternatively, we can visualize holding our flattened right hand out with the fingers aligned

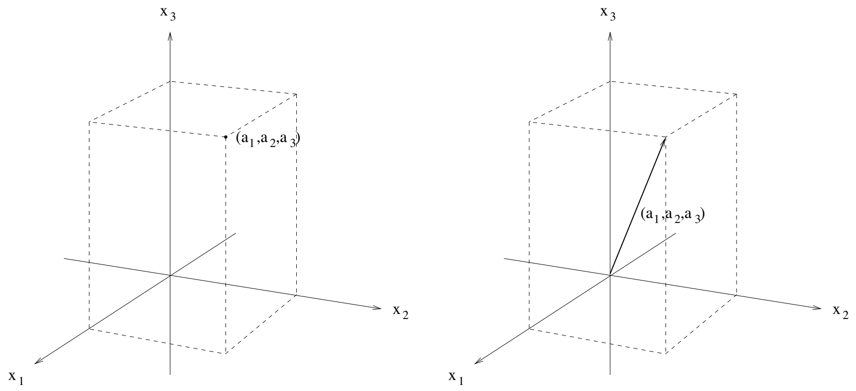


FIGURE 21.1. Coordinate system for $\mathbb{R}^3$

along the x_1 axis so that when we curl our fingers inward, they move towards the positive x_2 axis, and then our extended thumb will point along the positive x_3 axis.

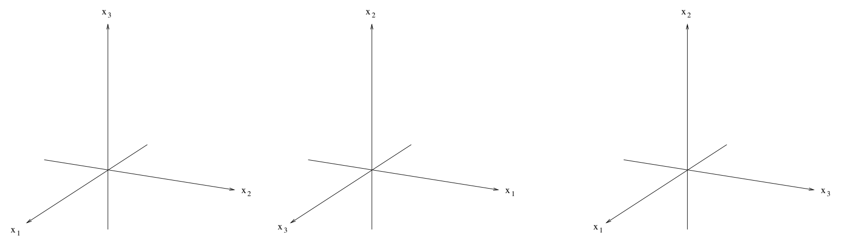


FIGURE 21.2. Two "right" coordinate systems and one "left", where the vertical coordinate of the view point is assumed positive, that is, the horizontal plane is seen from above. What happens if the vertical coordinate of the view point is assumed negative?

Having now chosen a right-oriented orthogonal coordinate system, we can assign three coordinates (a_1, a_2, a_3) to each point a in space using the same principle as in the case of the Euclidean plane, see Fig. 21.1. This way we can represent Euclidean 3d space as the set of all ordered 3-tuples $a = (a_1, a_2, a_3)$, where $a_i \in \mathbb{R}$ for $i = 1, 2, 3$, or as $\mathbb{R}^3$. Of course, we can choose different coordinate systems with different origin, coordinate directions and scaling of the coordinate axes. Below, we will come back to the topic of changing from one coordinate system to another.

21.2 Vector addition and multiplication by a scalar

Most of the notions and concepts of analytic geometry of the Euclidean plane represented by $\mathbb{R}^2$ extend naturally to Euclidean 3d space represented by $\mathbb{R}^3$.

In particular, we can view an ordered 3-tuple $a = (a_1, a_2, a_3)$ either as a point in three dimensional space with coordinates a_1, a_2 and a_3 or as a vector/arrow with tail at the origin and head at the point (a_1, a_2, a_3), as illustrated in Fig. 21.1.

We define the sum $a+b$ of two vectors $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$ in $\mathbb{R}^3$ by componentwise addition,

$$a + b = (a_1 + b_1, a_2 + b_2, a_3 + b_3),$$

and multiplication of a vector $a = (a_1, a_2, a_3)$ by a real number λ by

$$\lambda\, a = (\lambda a_1, \lambda a_2, \lambda a_3).$$

The *zero vector* is the vector $0 = (0, 0, 0)$. We also write $-a = (-1)a$ and $a - b = a + (-1)b$. The geometric interpretation of these definitions is analogous to that in $\mathbb{R}^2$. For example, two non-zero vectors a and b in $\mathbb{R}^3$ are *parallel* if $b = \lambda a$ for some non-zero real number λ. The usual rules hold, so vector addition is *commutative*, $a + b = b + a$, and *associative*, $(a + b) + c = a + (b + c)$. Further, $\lambda(a + b) = \lambda a + \lambda b$ and $\kappa(\lambda a) = (\kappa\lambda)a$ for vectors a and b and real numbers λ and κ.

The standard basis vectors in $\mathbb{R}^3$ are $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$ and $e_3 = (0, 0, 1)$.

21.3 Scalar product and norm

The standard *scalar product* $a \cdot b$ of two vectors $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$ in $\mathbb{R}^3$ is defined by

$$a \cdot b = \sum_{i=1}^{3} a_i b_i = a_1 b_1 + a_2 b_2 + a_3 b_3. \tag{21.1}$$

The scalar product in $\mathbb{R}^3$ has the same properties as its cousin in $\mathbb{R}^2$, so it is bilinear, symmetric, and positive definite.= We say that two vectors a and b are *orthogonal* if $a \cdot b = 0$.

The Euclidean *length* or *norm* $|a|$ of a vector $a = (a_1, a_2, a_3)$ is defined by

$$|a| = (a \cdot a)^{\frac{1}{2}} = \Big(\sum_{i=1}^{3} a_i^2\Big)^{\frac{1}{2}}, \tag{21.2}$$

which expresses Pythagoras theorem in 3d, and which we may obtain by using the usual 2d Pythagoras theorem twice. The distance $|a-b|$ between two points $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$ is equal to

$$|a-b| = \Big(\sum_{i=1}^{3}(a_i - b_i)^2\Big)^{1/2}.$$

Cauchy's inequality states that for any vectors a and b in $\mathbb{R}^3$,

$$|a \cdot b| \leq |a|\,|b|. \tag{21.3}$$

We give a proof of Cauchy's inequality in Chapter Analytic Geometry in $\mathbb{R}^n$ below. We note that Cauchy's inequality in $\mathbb{R}^2$ follows directly from the fact that $a \cdot b = |a|\,|b|\,\cos(\theta)$, where θ is the angle between a and b.

21.4 Projection of a vector onto a vector

Let a be a given non-zero vector in $\mathbb{R}^3$. We define the projection $Pb = P_a(b)$ of a vector b in $\mathbb{R}^3$ onto the vector a by the formula

$$Pb = P_a(b) = \frac{a \cdot b}{a \cdot a}\,a = \frac{a \cdot b}{|a|^2}\,a. \tag{21.4}$$

This is a direct generalization of the corresponding formula in $\mathbb{R}^2$ based on the principles that Pb is parallel to a and $b - Pb$ is orthogonal to a as illustrated in Fig. 21.3, that is

$$Pb = \lambda a \quad \text{for some } \lambda \in \mathbb{R} \quad \text{and} \quad (b - Pb) \cdot a = 0.$$

This gives the formula (21.4) with $\lambda = \frac{a \cdot b}{|a|^2}$.

The transformation $P : \mathbb{R}^3 \to \mathbb{R}^3$ is linear, that is for any b and $c \in \mathbb{R}^3$ and $\lambda \in \mathbb{R}$,

$$P(b + c) = Pb + Pc, \quad P(\lambda b) = \lambda Pb,$$

and $PP = P$.

21.5 The angle between two vectors in $\mathbb{R}^3$

We define the angle θ between non-zero vectors a and b in $\mathbb{R}^3$ by

$$\cos(\theta) = \frac{a \cdot b}{|a|\,|b|}, \tag{21.5}$$

where we may assume that $0 \leq \theta \leq 180°$. By Cauchy's inequality (21.3), $|a \cdot b| \leq |a|\,|b|$. Thus, there is an angle θ satisfying (21.5) that is uniquely defined if we require $0 \leq \theta \leq 180°$. We may write (21.5) in the form

$$a \cdot b = |a||b|\cos(\theta), \tag{21.6}$$

where θ is the angle between a and b. This evidently extends the corresponding result in $\mathbb{R}^2$.

We define the angle θ between two vectors a and b via the scalar product $a \cdot b$ in (21.5), which we may view as an *algebraic* definition. Of course, we would like to see that this definition coincides with a usual *geometric* definition. If a and b both lie in the $x_1 - x_2$-plane, then we know from the Chapter Analytic geometry in $\mathbb{R}^2$ that the two definitions coincide. We shall see below that the scalar product $a \cdot b$ is invariant (does not change) under rotation of the coordinate system, which means that given any two vectors a and b, we can rotate the coordinate system so make a and b lie in the $x_1 - x_2$-plane. We conclude that the algebraic definition (21.5) of angle between two vectors and the usual geometric definition coincide. In particular, two non-zero vectors are geometrically orthogonal in the sense that the geometric angle θ between the vectors satisfies $\cos(\theta) = 0$ if and only if $a \cdot b = |a||b|\cos(\theta) = 0$.

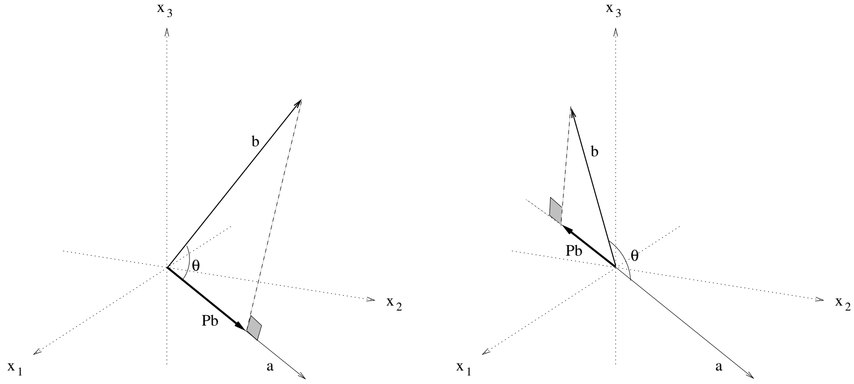


FIGURE 21.3. Projection Pb of a vector b onto a vector a.

21.6 Vector product

We now define the *vector product* $a \times b$ of two vectors $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$ in $\mathbb{R}^3$ by the formula

$$a \times b = (a_2 b_3 - a_3 b_2, a_3 b_1 - a_1 b_3, a_1 b_2 - a_2 b_1). \tag{21.7}$$

We note that the vector product $a \times b$ of two vectors a and b in $\mathbb{R}^3$ is itself a vector in $\mathbb{R}^3$. In other words, with $f(a, b) = a \times b$, $f : \mathbb{R}^3 \to \mathbb{R}^3$. We also refer to the vector product as the *cross product*, because of the notation.

EXAMPLE 21.1. If $a = (3, 2, 1)$ and $b = (4, 5, 6)$, then $a \times b = (12 - 5, 4 - 18, 15 - 8) = (7, -14, 7)$.

Note that there is also the trivial, componentwise "vector product" defined by (using *MATLAB*© 's notation) $a.{*}b = (a_1 b_1, a_2 b_2, a_3 b_3)$. The vector product defined above, however, is something quite different!

The formula for the vector product may seem a bit strange (and complicated), and we shall now see how it arises. We start by noting that the expression $a_1 b_2 - a_2 b_1$ appearing in (21.7) is the vector product of the vectors (a_1, a_2) and (b_1, b_2) in $\mathbb{R}^2$, so there appears to be some pattern at least.

We may directly check that the vector product $a \times b$ is linear in both a and b, that is

$$a \times (b + c) = a \times b + a \times c, \quad (a + b) \times c = a \times c + b \times c, \tag{21.8a}$$

$$(\lambda a) \times b = \lambda\, a \times b, \quad a \times (\lambda b) = \lambda\, a \times b, \tag{21.8b}$$

where the products $\times$ should be computed first unless something else is indicated by parentheses. This follows directly from the fact that the components of $a \times b$ depend linearly on the components of a and b.

Since the vector product $a \times b$ is linear in both a and b, we say that $a \times b$ is *bilinear*. We also see that the vector product $a \times b$ is *anti-symmetric* in the sense that

$$a \times b = -\, b \times a. \tag{21.9}$$

Thus, the vector product $a \times b$ is bilinear and antisymmetric and moreover it turns out that these two properties determine the vector product up to a constant just as in in $\mathbb{R}^2$.

For the vector products of the basis vectors e_i, we have (check this!)

$$e_i \times e_i = 0,\ i = 1, 2, 3, \tag{21.10a}$$

$$e_1 \times e_2 = e_3, \quad e_2 \times e_3 = e_1, \quad e_3 \times e_1 = e_2, \tag{21.10b}$$

$$e_2 \times e_1 = -e_3, \quad e_3 \times e_2 = -e_1, \quad e_1 \times e_3 = -e_2. \tag{21.10c}$$

We see that $e_1 \times e_2 = e_3$ is orthogonal to both e_1 and e_2. Similarly, $e_2 \times e_3 = e_1$ is orthogonal to both e_2 and e_3, and $e_1 \times e_3 = -e_2$ is orthogonal to both e_1 and e_3.

This pattern generalizes. In fact, for any two non-zero vectors a and b, the vector $a \times b$ is orthogonal to both a and b since

$$a \cdot (a \times b) = a_1(a_2 b_3 - a_3 b_2) + a_2(a_3 b_1 - a_1 b_3) + a_3(a_1 b_2 - a_2 b_1) = 0, \tag{21.11}$$

and similarly $b \cdot (a \times b) = 0$.

We may compute the vector product of two arbitrary vectors $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$ by using linearity combined with (21.10) as follows,

$$\begin{aligned} a \times b &= (a_1e_1 + a_2e_2 + a_3e_3) \times (b_1e_1 + b_2e_2 + b_3e_3) \\ &= a_1b_2\, e_1 \times e_2 + a_2b_1\, e_2 \times e_1 \\ &\qquad + a_1b_3\, e_1 \times e_3 + a_3b_1\, e_3 \times e_1 \\ &\qquad\qquad + a_2b_3\, e_2 \times e_3 + a_3b_2\, e_3 \times e_2 \\ &= (a_1b_2 - a_2b_1)e_3 + (a_3b_1 - a_1b_3)e_2 + (a_2b_3 - a_3b_2)e_1 \\ &= (a_2b_3 - a_3b_2, a_3b_1 - a_1b_3, a_1b_2 - a_2b_1), \end{aligned}$$

which conforms with (21.7).

21.7 Geometric interpretation of the vector product

We shall now make a geometric interpretation of the vector product $a \times b$ of two vectors a and b in $\mathbb{R}^3$.

We start by assuming that $a = (a_1, a_2, 0)$ and $b = (b_1, b_2, 0)$ are two non-zero vectors in the plane defined by the x_1 and x_2 axes. The vector $a \times b = (0, 0, a_1b_2 - a_2b_1)$ is clearly orthogonal to both a and b, and recalling the basic result (20.21) for the vector product in $\mathbb{R}^2$, we have

$$|a \times b| = |a||b||\sin(\theta)|, \tag{21.12}$$

where θ is the angle between a and b.

We shall now prove that this result generalizes to arbitrary vectors $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$ in $\mathbb{R}^3$. First, the fact that $a \times b$ is orthogonal to both a and b was proved in the previous section. Secondly, we note that multiplying the trigonometric identity $\sin^2(\theta) = 1 - \cos^2(\theta)$ by $|a|^2\,|b|^2$ and using (21.6), we obtain

$$|a|^2|b|^2 \sin^2(\theta) = |a|^2|b|^2 - (a \cdot b)^2. \tag{21.13}$$

Finally, a direct (but somewhat lengthy) computation shows that

$$|a \times b|^2 = |a|^2|b|^2 - (a \cdot b)^2,$$

which proves (21.12). We summarize in the following theorem.

Theorem 21.1 *The vector product $a \times b$ of two non-zero vectors a and b in $\mathbb{R}^3$ is orthogonal to both a and b and $|a \times b| = |a||b||\sin(\theta)|$, where θ is the angle between a and b. In particular, a and b are parallel if and only if $a \times b = 0$.*

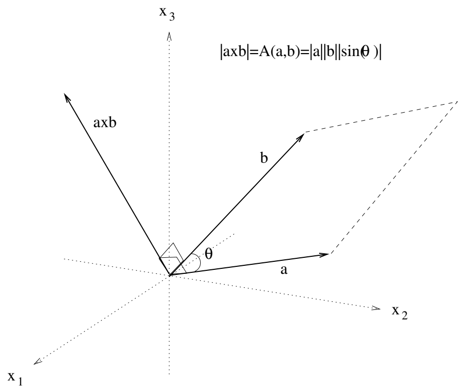


FIGURE 21.4. Geometric interpretation of the vector product.

We can make the theorem more precise by adding the following sign rule: The vector $a \times b$ is pointing in the direction of a standard screw turning the vector a into the vector b the shortest route.

21.8 Connection of vector products in $\mathbb{R}^2$ and $\mathbb{R}^3$

We note that if $a = (a_1, a_2, 0)$ and $b = (b_1, b_2, 0)$, then

$$a \times b = (0, 0, a_1 b_2 - a_2 b_1). \tag{21.14}$$

The previous formula $a \times b = a_1 b_2 - a_2 b_1$ for $a = (a_1, a_2)$ and $b = (b_1, b_2)$ in $\mathbb{R}^2$ may thus be viewed as a short-hand for the formula $a \times b = (0, 0, a_1 b_2 - a_2 b_1)$ for $a = (a_1, a_2, 0)$ and $b = (b_1, b_2, 0)$, with $a_1 b_2 - a_2 b_1$ being the third coordinate of $a \times b$ in $\mathbb{R}^3$. We note the relation of the sign conventions in $\mathbb{R}^2$ and $\mathbb{R}^3$: If $a_1 b_2 - a_2 b_1 \geq 0$, then turning a screw into the positive x_3-direction should turn a into b the shortest route. This corresponds to turning a into b counter-clockwise and to the angle θ between a and b satisfying $\sin(\theta) \geq 0$.

21.9 Volume of parallelepiped spanned by three vectors

Consider the parallelepiped spanned by three vectors a, b and c, according to Fig. 21.5.

We seek a formula for the *volume* $V(a, b, c)$ of the parallelepiped. We recall that the volume $V(a, b, c)$ is equal to the area $A(a, b)$ of the *base* spanned by the vectors a and b times the height h, which is the length of

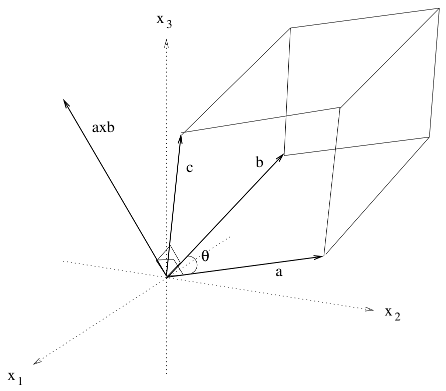


FIGURE 21.5. Parallelepiped spanned by three vectors

the projection of c onto a vector that is orthogonal to the plane formed by a and b. Since $a \times b$ is orthogonal to both a and b, the height h is equal to the length of the projection of c onto $a \times b$. From (21.12) and (21.4), we know that

$$A(a,b) = |a \times b|, \quad h = \frac{|c \cdot (a \times b)|}{|a \times b|},$$

and thus

$$V(a,b,c) = |c \cdot (a \times b)|. \tag{21.15}$$

Clearly, we may also compute the volume $V(a,b,c)$ by considering b and c as forming the base, or likewise the vectors a and c forming the base. Thus,

$$V(a,b,c) = |a \cdot (b \times c)| = |b \cdot (a \times c)| = |c \cdot (a \times b)|. \tag{21.16}$$

EXAMPLE 21.2. The volume $V(a,b,c)$ of the parallelepiped spanned by $a = (1,2,3)$, $b = (3,2,1)$ and $c = (1,3,2)$ is equal to $a \cdot (b \times c) = (1,2,3) \cdot (1,-5,7) = 12$.

21.10 The triple product $a \cdot b \times c$

The expression $a \cdot (b \times c)$ occurs in the formulas (21.15) and (21.16). This is called the *triple product* of the three vectors a, b and c. We usually write the triple product without the parenthesis following the convention that the vector product $\times$ is performed first. In fact, the alternative interpretation $(a \cdot b) \times c$ does not make sense since $a \cdot b$ is a scalar and the vector product $\times$ requires vector factors!

The following properties of the triple product can be readily verified by direct application of the definition of the scalar and vector products,

$$a \cdot b \times c = c \cdot a \times b = b \cdot c \times a,$$
$$a \cdot b \times c = -a \cdot c \times b = -b \cdot a \times c = -c \cdot b \times a.$$

To remember these formulas, we note that if two of the vectors change place then the sign changes, while if all three vectors are cyclically permuted (for example the order a, b, c is replaced by c, a, b or b, c, a), then the sign is unchanged.

Using the triple product $a \cdot b \times c$, we can express the geometric quantity of the volume $V(a, b, c)$ of the parallelepiped spanned by a, b and c in the concise algebraic form,

$$V(a, b, c) = |a \cdot b \times c|. \tag{21.17}$$

We shall use this formula many times below. Note, we later prove that the volume of a parallelepiped can be computed as the area of the base times the height using Calculus below.

21.11 An explicit formula for the volume $V(a_1, a_2, a_3)$ spanned by three vectors a_1, a_2 and a_3

Let $a_1 = (a_{11}, a_{12}, a_{13})$, $a_2 = (a_{21}, a_{22}, a_{23})$ and $a_3 = (a_{31}, a_{32}, a_{33})$ be three vectors in $\mathbb{R}^3$. Note that here a_1 is a vector in $\mathbb{R}^3$ with $a_1 = (a_{11}, a_{12}, a_{13})$, et cetera. We may think of forming the 3×3 *matrix* $A = (a_{ij})$ with the rows corresponding to the coordinates of a_1, a_2 and a_3,

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

We will come back to 3×3 matrices below. Here, we just use the matrix to express the coordinates of the vectors a_1, a_2 and a_3 in handy form.

We give an explicit formula for the volume $V(a_1, a_2, a_3)$ spanned by three vectors a_1, a_2 and a_3. By direct computation starting with (21.17),

$$\begin{aligned} \pm V(a_1, a_2, a_3) &= a_1 \cdot a_2 \times a_3 \\ &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) \\ &\quad + a_{13}(a_{21}a_{32} - a_{22}a_{31}). \end{aligned} \tag{21.18}$$

We note that $V(a_1, a_2, a_3)$ is a sum of terms, each term consisting of the product of three factors $a_{ij}a_{kl}a_{mn}$ with certain indices ij, kl and mn. If we examine the indices occurring in each term, we see that the sequence of row indices ikm (first indices) is always $\{1,2,3\}$, while the sequence of column indices jln (second indices) corresponds to a *permutation* of the sequence $\{1,2,3\}$, that is the numbers 1, 2 and 3 occur in some order. Thus, all terms have the form

$$a_{1j_1}a_{2j_2}a_{3j_3} \tag{21.19}$$

with $\{j_1,j_2,j_3\}$ being a permutation of $\{1,2,3\}$. The sign of the terms change with the permutation. By inspection we can detect the following pattern: if the permutation can be brought to the order $\{1,2,3\}$ with an even number of *transpositions*, each transposition consisting of interchanging two indices, then the sign is $+$, and with an uneven number of transpositions the sign is $-$. For example, the permutation of second indices in the term $a_{11}a_{23}a_{32}$ is $\{1,3,2\}$, which is uneven since one transposition brings it back to $\{1,2,3\}$, and thus this term has a negative sign. Another example: the permutation in the term $a_{12}a_{23}a_{31}$ is $\{2,3,1\}$ is even since it results from the following two transpositions $\{2,1,3\}$ and $\{1,2,3\}$.

We have now developed a technique for computing volumes that we will generalize to $\mathbb{R}^n$ below. This will lead to *determinants*. We will see that the formula (42.54) states that the signed volume $\pm V(a_1,a_2,a_3)$ is equal to the determinant of the 3×3 matrix $A=(a_{ij})$.

21.12 Lines in $\mathbb{R}^3$

Let a be a given non-zero vector in $\mathbb{R}^3$ and let $\hat{x}$ be a given point in $\mathbb{R}^3$. The points x in $\mathbb{R}^3$ of the form

$$x=\hat{x}+sa,$$

where s varies over $\mathbb{R}$, form a *line* in $\mathbb{R}^3$ through the point $\hat{x}$ with direction a, see Fig. 21.6. If $\hat{x}=0$, then the line passes through the origin.

EXAMPLE 21.3. The line through $(1,2,3)$ in the direction $(4,5,6)$ is given by

$$x=(1,2,3)+s(4,5,6)=(1+4s,2+5s,3+6s)\quad s\in\mathbb{R}.$$

The line through $(1,2,3)$ and $(3,1,2)$ has the direction $(3,1,2)-(1,2,3)=(2,-1,-1)$, and is thus given by $x=(1,2,3)+s(2,-1,-1)$.

Note that by choosing other vectors to represent the direction of the line, it may also be represented, for example, as $x=(1,2,3)+\hat{s}(-2,1,1)$ or $x=(1,2,3)+\tilde{s}(6,-3,-3)$. Also, the "point of departure" on the line, corresponding to $s=0$, can be chosen arbitrarily on the line of course. For example, the point $(1,2,3)$ could be replaced by $(-1,3,4)$ which is another point on the line.

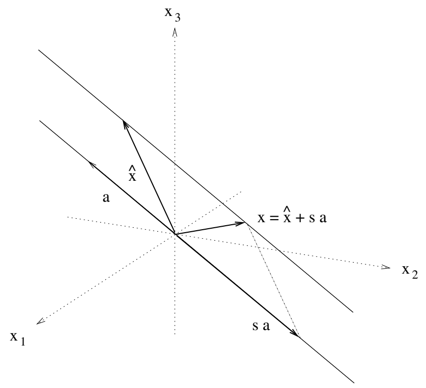


FIGURE 21.6. Line in $\mathbb{R}^3$ of the form $x = \hat{x} + sa$.

21.13 Projection of a point onto a line

Let $x = \hat{x} + sa$ be a line in $\mathbb{R}^3$ through $\hat{x}$ with direction $a \in \mathbb{R}^3$. We seek the *projection* Pb of a given *point* $b \in \mathbb{R}^3$ onto the line, that is we seek $Pb \in \mathbb{R}^3$ with the property that (i) $Pb = \hat{x} + sa$ for some $s \in \mathbb{R}$, and (ii) $(b - Pb) \cdot a = 0$. Note that we here view b to be a point rather than a vector. Inserting (i) into (ii) gives the following equation in s: $(b - \hat{x} - sa) \cdot a = 0$, from which we conclude that $s = \frac{b \cdot a - \hat{x} \cdot a}{|a|^2}$, and thus

$$Pb = \hat{x} + \frac{b \cdot a - \hat{x} \cdot a}{|a|^2} a \,. \tag{21.20}$$

If $\hat{x} = 0$, that is the line passes through the origin, then $Pb = \frac{b \cdot a}{|a|^2} a$ in conformity with the corresponding formula (20.9) in $\mathbb{R}^2$.

21.14 Planes in $\mathbb{R}^3$

Let a_1 and a_2 be two given non-zero non-parallel vectors in $\mathbb{R}^3$, that is $a_1 \times a_2 \neq 0$. The points x in $\mathbb{R}^3$ that can be expressed as

$$x = s_1 a_1 + s_2 a_2, \tag{21.21}$$

where s_1 and s_2 vary over $\mathbb{R}$, form a *plane* in $\mathbb{R}^3$ through the origin that is *spanned* by the two vectors a_1 and a_2. The points x in the plane are all the linear combinations $x = s_1 a_1 + s_2 a_2$ of the vectors a_1 and a_2 with coefficients s_1 and s_2 varying over $\mathbb{R}$, see Fig. 21.7. The vector $a_1 \times a_2$ is orthogonal to both a_1 and a_2 and therefore to all vectors x in the plane. Thus, the non-zero vector $n = a_1 \times a_2$ is a *normal* to the plane. The points x in the plane are characterized by the orthogonality relation

$$n \cdot x = 0. \tag{21.22}$$

We may thus describe the points x in the plane by the representation (21.21) or the equation (21.22). Note that (21.21) is a vector equation corresponding to 3 scalar equations, while (21.22) is a scalar equation. Eliminating the parameters s_1 and s_2 in the system (21.21), we obtain the scalar equation (21.22).

Let $\hat{x}$ be a given point in $\mathbb{R}^3$. The points x in $\mathbb{R}^3$ that can be expressed as

$$x = \hat{x} + s_1 a_1 + s_2 a_2, \tag{21.23}$$

where s_1 and s_2 vary over $\mathbb{R}$, form a *plane* in $\mathbb{R}^3$ through the point $\hat{x}$ that is parallel to the corresponding plane through the origin considered above, see Fig. 21.8.

If $x = \hat{x} + s_1 a_1 + s_2 a_2$ then $n \cdot x = n \cdot \hat{x}$, because $n \cdot a_i = 0$, $i = 1, 2$. Thus, we can describe the points x of the form $x = \hat{x} + s_1 a_1 + s_2 a_2$ alternatively as the vectors x satisfying

$$n \cdot x = n \cdot \hat{x}. \tag{21.24}$$

Again, we obtain the scalar equation (21.24) if we eliminate the parameters s_1 and s_2 in the system (21.23).

We summarize:

Theorem 21.2 *A plane in $\mathbb{R}^3$ through a point $\hat{x} \in \mathbb{R}^3$ with normal n can be expressed as the set of $x \in \mathbb{R}^3$ of the form $x = \hat{x} + s_1 a_1 + s_2 a_2$ with s_1 and s_2 varying over $\mathbb{R}$, where a_1 and a_2 are two vectors satisfying $n = a_1 \times a_2 \neq 0$. Alternatively, the plane can be described as the set of $x \in \mathbb{R}$ such that $n \cdot x = d$, where $d = n \cdot \hat{x}$.*

EXAMPLE 21.4. Consider the plane $x_1 + 2x_2 + 3x_3 = 4$, that is the plane $(1, 2, 3) \cdot (x_1, x_2, x_3) = 4$ with normal $n = (1, 2, 3)$. To write the points x in this plane on the form $x = \hat{x} + s_1 a_1 + s_2 a_2$, we first choose a point $\hat{x}$ in the plane, for example, $\hat{x} = (2, 1, 0)$ noting that $n \cdot \hat{x} = 4$. We next choose two non-parallel vectors a_1 and a_2 such that $n \cdot a_1 = 0$ and $n \cdot a_2 = 0$, for example $a_1 = (-2, 1, 0)$ and $a_2 = (-3, 0, 1)$. Alternatively, we choose one vector a_1 satisfying $n \cdot a_1 = 0$ and set $a_2 = n \times a_1$, which is a vector orthogonal to both n and a_1. To find a vector a_1 satisfying $a_1 \cdot n = 0$, we may choose an arbitrary non-zero vector m non-parallel to n and set $a_1 = m \times n$, for example $m = (0, 0, 1)$ giving $a_1 = (-2, 1, 0)$.

Conversely, given the plane $x = (2, 1, 0) + s_1(-2, 1, 0) + s_2(-3, 0, 1)$, that is $x = \hat{x} + s_1 a_1 + s_2 a_2$ with $\hat{x} = (2, 1, 0)$, $a_1 = (-2, 1, 0)$ and $a_2 = (-3, 0, 1)$, we obtain the equation $x_1 + 2x_2 + 3x_3 = 4$ simply by computing $n = a_1 \times a_2 = (1, 2, 3)$ and $n \cdot \hat{x} = (1, 2, 3) \cdot (2, 1, 0) = 4$, from which we obtain the following equation for the plane: $n \cdot x = (1, 2, 3) \cdot (x_1, x_2, x_3) = x_1 + 2x_2 + 3x_3 = n \cdot \hat{x} = 4$.

EXAMPLE 21.5. Consider the real-valued function $z = f(x, y) = ax + by + c$ of two real variables x and y, where a, b and c are real numbers.

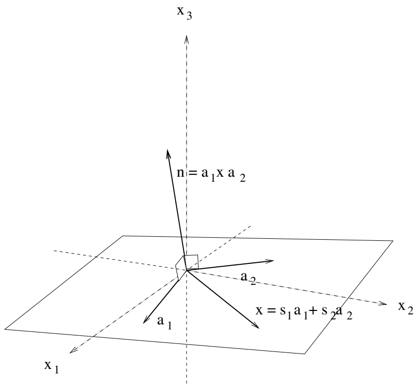


FIGURE 21.7. Plane through the origin spanned by b_1 and b_2, and with normal $n = b_1 \times b_2$.

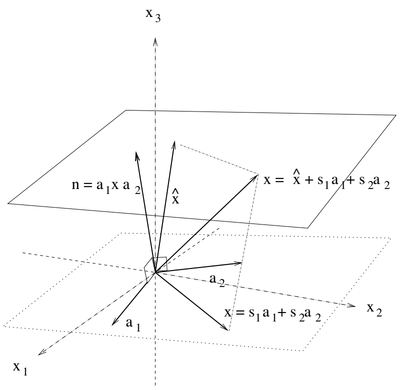


FIGURE 21.8. Plane through $\hat{x}$ with normal n defined by $n \cdot x = d = n \cdot \hat{x}$.

Setting $x_1 = x$, $x_2 = y$ and $x_3 = z$, we can express the graph of $z = f(x,y)$ as the plane $ax_1 + bx_2 - x_3 = -c$ in $\mathbb{R}^3$ with normal $(a, b, -1)$.

21.15 The intersection of a line and a plane

We seek the *intersection* of a line $x = \hat{x} + sa$ and a plane $n \cdot x = d$ that is the set of points x belonging to both the line and the plane, where $\hat{x}$, a, n and d are given. Inserting $x = \hat{x} + sa$ into $n \cdot x = d$, we obtain $n \cdot (\hat{x} + sa) = d$, that is $n \cdot \hat{x} + s\, n \cdot a = d$. This yields $s = (d - n \cdot \hat{x})/(n \cdot a)$ if $n \cdot a \neq 0$, and we find a unique point of intersection

$$x = \hat{x} + (d - n \cdot \hat{x})/(n \cdot a)\, a. \tag{21.25}$$

This formula has no meaning if $n \cdot a = 0$, that is if the line is parallel to the plane. In this case, there is no intersection point unless $\hat{x}$ happens to be a point in the plane and then the whole line is part of the plane.

EXAMPLE 21.6. The intersection of the plane $x_1 + 2x_2 + x_3 = 5$ and the line $x = (1, 0, 0) + s(1, 1, 1)$ is found by solving the equation $1 + s + 2s + s = 5$ giving $s = 1$ and thus the point of intersection is $(2, 1, 1)$.The plane $x_1 + 2x_2 + x_3 = 5$ and the line $x = (1, 0, 0) + s(2, -1, 0)$ has no point of intersection, because the equation $1 + 2s - 2s = 5$ has no solution. If instead we consider the plane $x_1 + 2x_2 + x_3 = 1$, we find that the entire line $x = (1, 0, 0) + s(2, -1, 0)$ lies in the plane, because $1 + 2s - 2s = 1$ for all real s.

21.16 Two intersecting planes define a line

Let $n_1 = (n_{11}, n_{12}, n_{13})$ and $n_2 = (n_{21}, n_{22}, n_{23})$ be two vectors in $\mathbb{R}^3$ and d_1 and d_2 two real numbers. The set of points $x \in \mathbb{R}^3$ that lie in both the plane $n_1 \cdot x = d_1$ and $n_2 \cdot x = d_2$ satisfy the system of two equations

$$\begin{aligned} n_1 \cdot x &= d_1, \\ n_2 \cdot x &= d_2. \end{aligned} \tag{21.26}$$

Intuition indicates that generally the points of intersection of the two planes should form a line in $\mathbb{R}^3$. Can we determine the formula of this line in the form $x = \hat{x} + sa$ with suitable vectors a and $\hat{x}$ in $\mathbb{R}^3$ and s varying over $\mathbb{R}$? Assuming first that $d_1 = d_2 = 0$, we seek a formula for the set of x such that $n_1 \cdot x = 0$ and $n_2 \cdot x = 0$, that is the set of x that are orthogonal to both n_1 and n_2. This leads to $a = n_1 \times n_2$ and expressing the solution x of the equations $n_1 \cdot x = 0$ and $n_2 \cdot x = 0$ as $x = s\, n_1 \times n_2$ with $s \in \mathbb{R}$. Of course it is natural to add in the assumption that $n_1 \times n_2 \neq 0$, that is that the two normals n_1 and n_2 are not parallel so that the two planes are not parallel.

Next, suppose that $(d_1, d_2) \neq (0,0)$. We see that if we can find one vector $\hat{x}$ such that $n_1 \cdot \hat{x} = d_1$ and $n_2 \cdot \hat{x} = d_2$, then we can write the solution x of (21.26) as

$$x = \hat{x} + s\, n_1 \times n_2, \quad s \in \mathbb{R}. \tag{21.27}$$

We now need to verify that we can indeed find $\hat{x}$ satisfying $n_1 \cdot \hat{x} = d_1$ and $n_2 \cdot \hat{x} = d_2$. That is, we need to find $\hat{x} \in \mathbb{R}^3$ satisfying the following system of two equations,

$$\begin{aligned} n_{11}\hat{x}_1 + n_{12}\hat{x}_2 + n_{13}\hat{x}_3 &= d_1, \\ n_{21}\hat{x}_1 + n_{22}\hat{x}_2 + n_{23}\hat{x}_3 &= d_2. \end{aligned}$$

Since $n_1 \times n_2 \neq 0$, some component of $n_1 \times n_2$ must be nonzero. If for example $n_{11}n_{22} - n_{12}n_{21} \neq 0$, corresponding to the third component of $n_1 \times n_2$ being non-zero, then we may choose $\hat{x}_3 = 0$. Then recalling the role of the condition $n_{11}n_{22} - n_{12}n_{21} \neq 0$ for a 2×2-system, we may solve uniquely for $\hat{x}_1$ and $\hat{x}_2$ in terms of d_1 and d_2 to get a desired $\hat{x}$. The argument is similar in case the second or first component of $n_1 \times n_2$ happens to be non-zero.

We summarize:

Theorem 21.3 *Two non-parallel planes $n_1 \cdot x = d_1$ and $n_2 \cdot x = d_2$ with normals n_1 and n_2 satisfying $n_1 \times n_2 \neq 0$, intersect along a straight line with direction $n_1 \times n_2$.*

EXAMPLE 21.7. The intersection of the two planes $x_1 + x_2 + x_3 = 2$ and $3x_1 + 2x_2 - x_3 = 1$ is given by $x = \hat{x} + sa$ with $a = (1,1,1) \times (3,2,-1) = (-3,4,-1)$ and $\hat{x} = (0,1,1)$.

21.17 Projection of a point onto a plane

Let $n \cdot x = d$ be a plane in $\mathbb{R}^3$ with normal n and b a point in $\mathbb{R}^3$. We seek the *projection* Pb of b onto the plane $n \cdot x = d$. It is natural to ask Pb to satisfy the following two conditions, see Fig. 21.9,

$$\begin{gathered} n \cdot Pb = d, \text{ that is } Pb \text{ is a point in the plane,} \\ b - Pb \text{ is parallel to the normal } n, \text{ that is } b - Pb = \lambda n \text{ for some } \lambda \in \mathbb{R}. \end{gathered}$$

We conclude that $Pb = b - \lambda n$ and the equation $n \cdot Pb = d$ thus gives $n \cdot (b - \lambda n) = d$. So, $\lambda = \frac{b \cdot n - d}{|n|^2}$ and thus

$$Pb = b - \frac{b \cdot n - d}{|n|^2}\, n. \tag{21.28}$$

If $d = 0$ so the plane $n \cdot x = d = 0$ passes through the origin, then

$$Pb = b - \frac{b \cdot n}{|n|^2} n. \tag{21.29}$$

If the plane is given in the form $x = \hat{x} + s_1 a_1 + s_2 a_2$ with a_1 and a_2 two given non-parallel vectors in $\mathbb{R}^3$, then we may alternatively compute the projection Pb of a point b onto the plane by seeking real numbers x_1 and x_2 so that $Pb = \hat{x} + x_1 a_1 + x_2 a_2$ and $(b - Pb) \cdot a_1 = (b - Pb) \cdot a_2 = 0$. This gives the system of equations

$$\begin{aligned} x_1 a_1 \cdot a_1 + x_2 a_2 \cdot a_1 &= b \cdot a_1 - \hat{x} \cdot a_1, \\ x_1 a_1 \cdot a_2 + x_2 a_2 \cdot a_2 &= b \cdot a_2 - \hat{x} \cdot a_2 \end{aligned} \tag{21.30}$$

in the two unknowns x_1 and x_2. To see that this system has a unique solution, we need to verify that $\hat{a}_{11}\hat{a}_{22} - \hat{a}_{12}\hat{a}_{21} \neq 0$, where $\hat{a}_{11} = a_1 \cdot a_1$, $\hat{a}_{22} = a_2 \cdot a_2$, $\hat{a}_{21} = a_2 \cdot a_1$ and $\hat{a}_{12} = a_1 \cdot a_2$. This follows from the fact that a_1 and a_2 are non-parallel, see Problem 21.24.

Example 21.8. The projection Pb of the point $b = (2, 2, 3)$ onto the plane defined by $x_1 + x_2 + x_3 = 1$ is given by $Pb = (2, 2, 3) - \frac{7-1}{3}(1, 1, 1) = (0, 0, 1)$.

Example 21.9. The projection Pb of the point $b = (2, 2, 3)$ onto the plane $x = (1, 0, 0) + s_1(1, 1, 1) + s_2(1, 2, 3)$ with normal $n = (1, 1, 1) \times (1, 2, 3) = (1, -2, 1)$ is given by $Pb = (2, 2, 3) - \frac{(2,2,3)\cdot(1,-2,1)}{6}(1, -2, 1) = (2, 2, 3) - \frac{1}{6}(1, -2, 1) = \frac{1}{6}(11, 14, 17)$.

21.18 Distance from a point to a plane

We say that the *distance* from a point b to a plane $n \cdot x = d$ is equal to $|b - Pb|$, where Pb is the projection of b onto the plane. According to the previous section, we have

$$|b - Pb| = \frac{|b \cdot n - d|}{|n|}.$$

Note that this distance is equal to the *shortest distance* between b and any point in the plane, see Fig. 21.9 and Problem 21.22.

Example 21.10. The distance from the point $(2, 2, 3)$ to the plane $x_1 + x_2 + x_3 = 1$ is equal to $\frac{|(2,2,3)\cdot(1,1,1)-1|}{\sqrt{3}} = 2\sqrt{3}$.

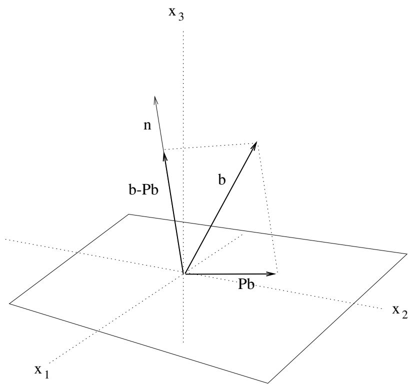


FIGURE 21.9. Projection of a point/vector onto a plane.

21.19 Rotation around a given vector

We now consider a more difficult problem. Let $a \in \mathbb{R}^3$ be a given vector and $\theta \in \mathbb{R}$ a given angle. We seek the transformation $R : \mathbb{R}^3 \to \mathbb{R}^3$ corresponding to rotation of an angle θ around the vector a. Recalling Section 20.21, the result $Rx = R(x)$ should satisfy the following properties,

$$\text{(i)} \quad |Rx - Px| = |x - Px|, \quad \text{(ii)} \quad (Rx - Px) \cdot (x - Px) = \cos(\theta)|x - Px|^2.$$

where $Px = P_a(x)$ is the projection of x onto a, see Fig. 21.10. We write

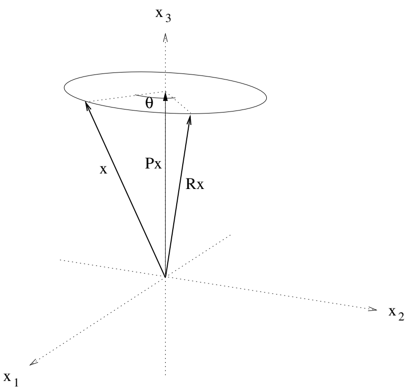


FIGURE 21.10. Rotation around $a = (0, 0, 1)$ a given angle θ.

$Rx - Px$ as $Rx - Px = \alpha(x - Px) + \beta\, a \times (x - Px)$ for real numbers α and β, noting that $Rx - Px$ is orthogonal to a and $a \times (x - Px)$ is orthogonal to both a and $(x - Px)$. Taking the scalar product with $(x - Px)$, we use (ii) to get $\alpha = \cos(\theta)$ and then use (i) to find $\beta = \frac{\sin(\theta)}{|a|}$ with a suitable orientation. Thus, we may express Rx in terms of the projection Px as

$$Rx = Px + \cos(\theta)(x - Px) + \frac{\sin(\theta)}{|a|} a \times (x - Px). \tag{21.31}$$

21.20 Lines and planes through the origin are subspaces of $\mathbb{R}^3$

Lines and planes in $\mathbb{R}^3$ through the origin are examples of *subspaces* of $\mathbb{R}^3$. The characteristic feature of a subspace is that the operations of vector addition and scalar multiplication does not lead outside the subspace. For example if x and y are two vectors in the plane through the origin with normal n satisfying $n \cdot x = 0$ and $n \cdot y = 0$, then $n \cdot (x + y) = 0$ and $n \cdot (\lambda x) = 0$ for any $\lambda \in \mathbb{R}$, so the vectors $x + y$ and λx also belong to the plane. On the other hand, if x and y belong to a plane not passing through the origin with normal n, so that $n \cdot x = d$ and $n \cdot y = d$ with d a nonzero constant, then $n \cdot (x + y) = 2d \neq d$, and thus $x + y$ does not lie in the plane. We conclude that lines and planes through the origin are subspaces of $\mathbb{R}^3$, but lines and planes not passing through the origin are not subspaces. The concept of subspace is very basic and we will meet this concept many times below.

We note that the equation $n \cdot x = 0$ defines a line in $\mathbb{R}^2$ and a plane in $\mathbb{R}^3$. The equation $n \cdot x = 0$ imposes a constraint on x that reduces the dimension by one, so in $\mathbb{R}^2$ we get a line and in $\mathbb{R}^3$ we get a plane.

21.21 Systems of 3 linear equations in 3 unknowns

Consider now the following system of 3 linear equations in 3 unknowns x_1, x_2 and x_3 (as did Leibniz already 1683):

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3, \end{aligned} \tag{21.32}$$

with coefficients a_{ij} and right hand side b_i, $i, j = 1, 2, 3$. We can write this system as the following vector equation in $\mathbb{R}^3$:

$$x_1 a_1 + x_2 a_2 + x_3 a_3 = b, \tag{21.33}$$

where $a_1 = (a_{11}, a_{21}, a_{31})$, $a_2 = (a_{12}, a_{22}, a_{32})$, $a_3 = (a_{13}, a_{23}, a_{33})$ and $b = (b_1, b_2, b_3)$ are vectors in $\mathbb{R}^3$, representing the given coefficients and the right hand side.

When is the system (21.32) uniquely solvable in $x = (x_1, x_2, x_3)$ for a given right hand side b? We shall see that the condition to guarantee unique solvability is

$$V(a_1, a_2, a_3) = |a_1 \cdot a_2 \times a_3| \neq 0, \qquad (21.34)$$

stating that the volume spanned by a_1, a_2 and a_3 is not zero.

We now argue that the condition $a_1 \cdot a_2 \times a_3 \neq 0$ is the right condition to guarantee the unique solvability of (21.32). We can do this by mimicking what we did in the case of a 2×2 system: Taking the scalar product of both sides of the vector equation $x_1 a_1 + x_2 a_2 + x_3 a_3 = b$ by successively $a_2 \times a_3$, $a_3 \times a_1$, and $a_2 \times a_3$, we get the following solution formula (recalling that $a_1 \cdot a_2 \times a_3 = a_2 \cdot a_3 \times a_1 = a_3 \cdot a_1 \times a_2$):

$$\begin{aligned} x_1 &= \frac{b \cdot a_2 \times a_3}{a_1 \cdot a_2 \times a_3}, \\ x_2 &= \frac{b \cdot a_3 \times a_1}{a_2 \cdot a_3 \times a_1} = \frac{a_1 \cdot b \times a_3}{a_1 \cdot a_2 \times a_3}, \\ x_3 &= \frac{b \cdot a_1 \times a_2}{a_3 \cdot a_1 \times a_2} = \frac{a_1 \cdot a_2 \times b}{a_1 \cdot a_2 \times a_3}, \end{aligned} \qquad (21.35)$$

where we used the facts that $a_i \cdot a_j \times a_k = 0$ if any two of the indices i, j and k are equal. The solution formula (21.35) shows that the system (21.32) has a unique solution if $a_1 \cdot a_2 \times a_3 \neq 0$.

Note the pattern of the solution formula (21.35), involving the common denominator $a_1 \cdot a_2 \times a_3$ and the numerator for x_i is obtained by replacing a_i by b. The solution formula (21.35) is also called *Cramer's rule*. We have proved the following basic result:

Theorem 21.4 *If* $a_1 \cdot a_2 \times a_3 \neq 0$*, then the system of equations* (21.32) *or the equivalent vector-equation* (21.33) *has a unique solution given by Cramer's rule* (21.35).

We repeat: $V(a_1, a_2, a_3) = a_1 \cdot a_2 \times a_3 \neq 0$ means that the three vectors a_1, a_2 and a_3 span a non-zero volume and thus point in three different directions (such that the plane spanned by any two of the vectors does not contain the third vector). If $V(a_1, a_2, a_3) \neq 0$, then we say that the set of three vectors $\{a_, a_2, a_3\}$ is *linearly independent*, or that the three vectors a_1, a_2 and a_3 are *linearly independent*.

21.22 Solving a 3×3-system by Gaussian elimination

We now describe an alternative to Cramer's rule for computing the solution to the 3×3-system of equations (21.32), using the famous method of *Gaussian elimination*. Assuming $a_{11} \neq 0$, we subtract the first equation multiplied by a_{21} from the second equation multiplied by a_{11}, and likewise subtract the first equation multiplied by by a_{31} from the third equation multiplied by a_{11}, to rewrite the system (21.32) in the form

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1, \\ (a_{22}a_{11} - a_{21}a_{12})x_2 + (a_{23}a_{11} - a_{21}a_{13})x_3 &= a_{11}b_2 - a_{21}b_1, \\ (a_{32}a_{11} - a_{31}a_{12})x_2 + (a_{33}a_{11} - a_{31}a_{13})x_3 &= a_{11}b_3 - a_{31}b_1, \end{aligned} \tag{21.36}$$

where the unknown x_1 has been *eliminiated* in the second and third equations. This system has the form

$$\begin{aligned} a_{11}x_1 + \quad a_{12}x_2 + a_{13}x_3 &= b_1, \\ \hat{a}_{22}x_2 + \hat{a}_{23}x_3 &= \hat{b}_2, \\ \hat{a}_{32}x_2 + \hat{a}_{33}x_3 &= \hat{b}_3, \end{aligned} \tag{21.37}$$

with modified coefficients $\hat{a}_{ij}$ and $\hat{b}_i$. We now proceed in the same way considering the 2×2-system in (x_2, x_3), and bring the system to the final triangular form

$$\begin{aligned} a_{11}x_1 + \quad a_{12}x_2 + a_{13}x_3 &= b_1, \\ \hat{a}_{22}x_2 + \hat{a}_{23}x_3 &= \hat{b}_2, \\ \tilde{a}_{33}x_3 &= \tilde{b}_3, \end{aligned} \tag{21.38}$$

with modified coefficients in the last equation. We can now solve the third equation for x_3, then insert the resulting value of x_3 into the second equation and solve for x_2 and finally insert x_3 and x_2 into the first equation to solve for x_1.

EXAMPLE 21.11. We give an example of Gaussian elimination: Consider the system

$$\begin{aligned} x_1 \quad +2x_2 + 3x_3 &= 6, \\ 2x_1 \quad +3x_2 + 4x_3 &= 9, \\ 3x_1 \quad +4x_2 + 6x_3 &= 13. \end{aligned}$$

Subtracting the first equation multiplied by 2 from the second and the first equation multiplied by 3 from the third equation, we get the system

$$\begin{aligned} x_1 \quad +2x_2 + 3x_3 &= 6, \\ -x_2 - 2x_3 &= -3, \\ -2x_2 - 3x_3 &= -5. \end{aligned}$$

Subtracting now the second equation multiplied by 2 from the third equation, we get

$$\begin{aligned} x_1 + 2x_2 + \quad 3x_3 &= 6, \\ -x_2 - \quad 2x_3 &= -3, \\ x_3 &= 1, \end{aligned}$$

from which we find $x_3 = 1$ and then from the second equation $x_2 = 1$ and finally from the first equation $x_1 = 1$.

21.23 3×3 matrices: sum, product and transpose

We can directly generalize the notion of a 2×2 matrix as follows: We say that the quadratic array

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

is a 3×3 *matrix* $A = (a_{ij})$ with elements a_{ij}, $i, j = 1, 2, 3$, and with i being the *row index* and j the *column index*.

Of course, we can also generalize the notion of a 2-row (or 1×2 matrix) and a 2-column vector (or 2×1 matrix). Each row of A, the first row being $(a_{11}\ a_{12}\ a_{13})$, can thus be viewed as a 3-row vector (or 1×3 matrix), and each column of A, the first column being

$$\begin{pmatrix} a_{11} \\ a_{21} \\ a_{31} \end{pmatrix}$$

as a 3-column vector (or 3×1 matrix). We can thus view a 3×3 matrix to consist of three 3-row vectors or three 3-column vectors.

Let $A = (a_{ij})$ and $B = (b_{ij})$ be two 3×3 matrices. We define the sum $C = A + B$ to be the matrix $C = (c_{ij})$ with elements $c_{ij} = a_{ij} + b_{ij}$ for $i, j = 1, 2, 3$. In other words, we add two matrices element by element.

Given a 3×3 matrix $A = (a_{ij})$ and a real number λ, we define the matrix $C = \lambda A$ as the matrix with elements $c_{ij} = \lambda a_{ij}$. In other words, all elements a_{ij} are multiplied by λ.

Given two 3×3 matrices $A = (a_{ij})$ and $B = (b_{ij})$, we define the product $C = AB$ as the 3×3 matrix with elements c_{ij} given by

$$c_{ij} = \sum_{k=1}^{3} a_{ik} b_{kj} \quad i, j = 1, 2, 3. \tag{21.39}$$

Matrix multiplication is *associative* so that $(AB)C = A(BC)$ for matrices A, B and C, see Problem 21.10. The matrix product is however not *commutative* in general, that is there are matrices A and B such that $AB \neq BA$, see Problem 21.11.

Given a 3×3 matrix $A = (a_{ij})$, we define the *transpose* of A denoted by $A^\top$ as the matrix $C = A^\top$ with elements $c_{ij} = a_{ji}$, $i, j = 1, 2, 3$. In other words, the rows of A are the columns of $A^\top$ and vice versa. By definition $(A^\top)^\top = A$. Transposing twice brings back the original matrix.

We can directly check the validity of the following rules for computing with the transpose:

$$(A + B)^\top = A^\top + B^\top, \quad (\lambda A)^\top = \lambda A^\top,$$
$$(AB)^\top = B^\top A^\top.$$

Similarly, the transpose of a 3-column vector is the 3-row vector with the same elements. Vice versa, if we consider the 3×1 matrix

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

to be a 3-column vector, then the transpose $x^\top$ is the corresponding 3-row vector $(x_1\ x_2\ x_3)$. We define the product of a 1×3 matrix (3-row vector) $x^\top$ with a 3×1 matrix (3-column vector) y in the natural way as follows:

$$x^\top y = \begin{pmatrix} x_1 & x_2 & x_3 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = x_1 y_1 + x_2 y_2 + x_3 y_3 = x \cdot y,$$

where we noted the connection to the scalar product of 3-vectors. We thus make the fundamental observation that multiplication of a 1×3 matrix (3-row vector) with a 3×1 matrix (3-column vector) is the same as scalar multiplication of the corresponding 3-vectors. We can then express the element c_{ij} of the product $C = AB$ according to (21.39) as the scalar product of row i of A with column j of B,

$$c_{ij} = \begin{pmatrix} a_{i1} & a_{i2} & a_{i3} \end{pmatrix} \begin{pmatrix} b_{1j} \\ b_{2j} \\ b_{3j} \end{pmatrix} = \sum_{k=1}^{3} a_{ik} b_{kj}.$$

We note that

$$|x|^2 = x \cdot x = x^\top x,$$

where we interpret x both as an ordered triple and as a 3-column vector.

The 3×3 matrix

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

is called the 3×3 *identity matrix* and is denoted by I. We have $IA = A$ and $AI = A$ for any 3×3 matrix A.

If $A = (a_{ij})$ is a 3×3 matrix and $x = (x_i)$ is a 3×1 matrix with elements x_i, then the product Ax is the 3×1 matrix with elements

$$\sum_{k=1}^{3} a_{ik} x_k \quad i = 1, 2, 3.$$

The linear system of equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1, \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2, \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3, \end{aligned}$$

can be written in matrix form as

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix},$$

that is

$$Ax = b,$$

with $A = (a_{ij})$ and $x = (x_i)$ and $b = (b_i)$.

21.24 Two different ways of viewing a system of linear equations

We may view a 3×3 matrix $A = (a_{ij})$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

as being formed by three column-vectors $a_1 = (a_{11}, a_{21}, a_{31})$, $a_2 = (a_{12}, a_{22}, a_{32})$, $a_3 = (a_{13}, a_{23}, a_{33})$, or by three row-vectors $\hat{a}_1 = (a_{11}, a_{12}, a_{13})$, $\hat{a}_2 = (a_{21}, a_{22}, a_{23})$, $\hat{a}_3 = (a_{31}, a_{32}, a_{33})$. Accordingly, we may view the system of equations

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix},$$

as a vector equation in the column vectors:

$$x_1 a_1 + x_2 a_2 + x_3 a_3 = b, \tag{21.40}$$

or as a system of 3 scalar equations:

$$\begin{aligned} \hat{a}_1 \cdot x &= b_1 \\ \hat{a}_2 \cdot x &= b_2 \\ \hat{a}_3 \cdot x &= b_3, \end{aligned} \tag{21.41}$$

where the rows $\hat{a}_i$ may be interpreted as normals to planes. We know from the discussion following (21.34) that (21.40) can be uniquely solved if $\pm V(a_1, a_2, a_3) = a_1 \cdot a_2 \times a_3 \neq 0$.

We also know from Theorem 21.3 that if $\hat{a}_2 \times \hat{a}_3 \neq 0$, then the set of $x \in \mathbb{R}^3$ satisfying the two last equations of (21.41) forms a line with direction $\hat{a}_2 \times \hat{a}_3$. If $\hat{a}_1$ is not orthogonal to $\hat{a}_2 \times \hat{a}_3$ then we expect this line to meet the plane given by the first equation of (21.41) at one point. Thus, if $\hat{a}_1 \cdot \hat{a}_2 \times \hat{a}_3 \neq 0$ then (21.41) should be uniquely solvable. This leads to the conjecture that $V(a_1, a_2, a_3) \neq 0$ if and only if $V(\hat{a}_1, \hat{a}_2, \hat{a}_3) \neq 0$. In fact, direct inspection from the formula (42.54) gives the more precise result,

Theorem 21.5 *If a_1, a_2 and a_3 are the vectors formed by the columns of a 3×3 matrix A, and $\hat{a}_1$, $\hat{a}_2$ and $\hat{a}_3$ are the vectors formed by the rows of A, then $V(a_1, a_2, a_3) = V(\hat{a}_1, \hat{a}_2, \hat{a}_3)$.*

21.25 Non-singular 3×3 matrices

Let A be a 3×3 matrix formed by three 3-column vectors a_1, a_2, and a_3. If $V(a_1, a_2, a_3) \neq 0$ then we say that A is *non-singular*, and if $V(a_1, a_2, a_3) = 0$ then we say that A is *singular*. From Section 21.21, we know that if A is non-singular then the matrix equation $Ax = b$ has a unique solution x for each $b \in \mathbb{R}^3$. Further, if A is singular then the three vectors a_1, a_2 and a_3 lie in the same plane and thus we can express one of the vectors as a linear combination of the other two. This implies that there is a non-zero vector $x = (x_1, x_2, x_3)$ such that $Ax = 0$. We sum up:

Theorem 21.6 *If A is a non-singular 3×3 matrix then the system of equations $Ax = b$ is uniquely solvable for any $b \in \mathbb{R}^3$. If A is singular then the system $Ax = 0$ has a non-zero solution x.*

21.26 The inverse of a matrix

Let A be a non-singular 3×3 matrix. Let $c_i \in \mathbb{R}^3$ be the solution to the equation $Ac_i = e_i$ for $i = 1, 2, 3$, where the e_i denote the standard basis vectors here interpreted as 3-column vectors. Let $C = (c_{ij})$ be the matrix with columns consisting of the vectors c_i. We then have $AC = I$, where I is the 3×3 identity matrix, because $Ac_i = e_i$. We call C the *inverse* of A and write $C = A^{-1}$ and note that A^{-1} is a 3×3 matrix such that

$$AA^{-1} = I, \tag{21.42}$$

that is multiplication of A by A^{-1} from the right gives the identity. We now want to prove that also $A^{-1}A = I$, that is that we get the identity also by multiplying A from the left by A^{-1}. To see this, we first note that A^{-1} must be non-singular, since if A^{-1} was singular then there would exist a non-zero vector x such that $A^{-1}x = 0$ and multiplying by A from the left would give $AA^{-1}x = 0$, contradicting the fact that by (21.42) $AA^{-1}x = Ix = x \neq 0$. Multiplying $AA^{-1} = I$ with A^{-1} from the left, we get $A^{-1}AA^{-1} = A^{-1}$, from which we conclude that $A^{-1}A = I$ by multiplying from the right with the inverse of A^{-1}, which we know exists since A^{-1} is non-singular.

We note that $(A^{-1})^{-1} = A$, which is a restatement of $A^{-1}A = I$, and that

$$(AB)^{-1} = B^{-1}A^{-1}$$

since $B^{-1}A^{-1}AB = B^{-1}B = I$. We summarize:

Theorem 21.7 *If A is a 3×3 non-singular matrix, then the inverse 3×3 matrix A^{-1} exists, and $AA^{-1} = A^{-1}A = I$. Further, $(AB)^{-1} = B^{-1}A^{-1}$.*

21.27 Different bases in $\mathbb{R}^3$

Let $\{a_1, a_2, a_3\}$ be a linearly independent set of three vectors in $\mathbb{R}^3$, that is assume that $V(a_1, a_2, a_3) \neq 0$. Theorem 21.4 implies that any given $b \in \mathbb{R}^3$ can be uniquely expressed as a linear combination of $\{a_1, a_2, a_3\}$,

$$b = x_1 a_1 + x_2 a_2 + x_3 a_3, \tag{21.43}$$

or in matrix language

$$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \qquad \text{or} \qquad b = Ax,$$

where the columns of the matrix $A = (a_{ij})$ are formed by the vectors $a_1 = (a_{11}, a_{21}, a_{31})$, $a_2 = (a_{12}, a_{22}, a_{32})$, $a_3 = (a_{13}, a_{23}, a_{33})$. Since $V(a_1, a_2, a_3) \neq 0$, the system of equations $Ax = b$ has a unique solution $x \in \mathbb{R}^3$ for any given $b \in \mathbb{R}^3$, and thus any $b \in \mathbb{R}^3$ can be expressed uniquely as a linear combination $b = x_1 a_1 + x_2 a_2 + x_3 a_3$ of the set of vectors $\{a_1, a_2, a_3\}$ with the coefficients (x_1, x_2, x_3). This means that $\{a_1, a_2, a_3\}$ is a *basis* for $\mathbb{R}^3$ and we say that (x_1, x_2, x_3) are the *coordinates* of b with respect to the basis $\{a_1, a_2, a_3\}$. The connection between the coordinates (b_1, b_2, b_3) of b in the standard basis and the coordinates x of b in the basis $\{a_1, a_2, a_3\}$ is given by $Ax = b$ or $x = A^{-1}b$.

21.28 Linearly independent set of vectors $\{a_1, a_2, a_3\}$ in $\mathbb{R}^3$

We say that a set of three vectors $\{a_1, a_2, a_3\}$ in $\mathbb{R}^3$ is *linearly independent* if $V(a_1, a_2, a_3) \neq 0$. We just saw that a linearly independent set $\{a_1, a_2, a_3\}$ of three vectors can be used as a basis in $\mathbb{R}^3$.

If the set $\{a_1, a_2, a_3\}$ is linearly independent then the system $Ax = 0$ in which the columns of the 3×3 matrix are formed by the coefficients of a_1, a_2 and a_3 has no other solution than $x = 0$.

Conversely, as a test of linear dependence we can use the following criterion: if $Ax = 0$ implies that $x = 0$, then $\{a_1, a_2, a_3\}$ is linearly independent and thus $V(a_1, a_2, a_3) \neq 0$.

We summarize:

Theorem 21.8 *A set $\{a_1, a_2, a_3\}$ of 3 vectors in $\mathbb{R}^3$ is linearly independent and can be used as a basis for $\mathbb{R}^3$ if $\pm V(a_1, a_2, a_3) = a_1 \cdot a_2 \times a_3 \neq 0$. A set*

$\{a_1, a_2, a_3\}$ of 3 vectors in $\mathbb{R}^3$ is linearly independent if and only if $Ax = 0$ implies that $x = 0$.

21.29 Orthogonal matrices

A 3×3 matrix Q satisfying $Q^\top Q = I$ is called an *orthogonal matrix*. An orthogonal matrix is non-singular with $Q^{-1} = Q^\top$ and thus also $QQ^\top = I$. An orthogonal matrix is thus characterized by the relation $Q^\top Q = QQ^\top = I$.

Let $q_i = (q_{1i}, q_{2i}, q_{3i})$ for $i = 1, 2, 3$, be the column vectors of Q, that is the row vectors of $Q^\top$. Stating that $Q^\top Q = I$ is the same as stating that

$$q_i \cdot q_j = 0 \quad \text{for } i \neq j, \quad \text{and } |q_i| = 1,$$

that is the columns of an orthogonal matrix Q are pairwise orthogonal and have length one.

EXAMPLE 21.12. The matrix

$$Q = \begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{21.44}$$

is orthogonal and corresponds to rotation of an angle θ around the x_3 axis.

21.30 Linear transformations versus matrices

Let $A = (a_{ij})$ be a 3×3 matrix. The mapping $x \to Ax$, that is the function $y = f(x) = Ax$, is a transformation from $\mathbb{R}^3$ to $\mathbb{R}^3$. This transformation is linear since $A(x+y) = Ax + Ay$ and $A(\lambda x) = \lambda Ax$ for $\lambda \in \mathbb{R}$. Thus, a 3×3 matrix A generates a linear transformation $f : \mathbb{R}^3 \to \mathbb{R}^3$ with $f(x) = Ax$.

Conversely to each linear transformation $f : \mathbb{R}^3 \to \mathbb{R}^3$, we can associate a matrix A with coefficients given by

$$a_{ij} = f_i(e_j)$$

where $f(x) = (f_1(x), f_2(x), f_3(x))$. The linearity of $f(x)$ implies

$$\begin{aligned} f(x) &= (f_1(\sum_{j=1}^3 x_j e_j), f_2(\sum_{j=1}^3 x_j e_j), f_3(\sum_{j=1}^3 x_j e_j))^\top \\ &= (\sum_{j=1}^3 f_1(e_j) x_j, \sum_{j=1}^3 f_2(e_j) x_j, \sum_{j=1}^3 f_3(e_j) x_j)^\top \\ &= (\sum_{j=1}^3 a_{1j} x_j, \sum_{j=1}^3 a_{2j} x_j, \sum_{j=1}^3 a_{3j} x_j)^\top = Ax, \end{aligned}$$

which shows that a linear transformation $f : \mathbb{R}^3 \to \mathbb{R}^3$ can be represented as $f(x) = Ax$ with the matrix $A = (a_{ij})$ with coefficients $a_{ij} = f_i(e_j)$.

EXAMPLE 21.13. The projection $Px = \frac{x \cdot a}{|a|^2} a$ onto a non-zero vector $a \in \mathbb{R}^3$ takes the matrix form

$$Px = \begin{pmatrix} \frac{a_1^2}{|a|^2} & \frac{a_1 a_2}{|a|^2} & \frac{a_1 a_3}{|a|^2} \\ \frac{a_2 a_1}{|a|^2} & \frac{a_2^2}{|a|^2} & \frac{a_2 a_3}{|a|^2} \\ \frac{a_3 a_1}{|a|^2} & \frac{a_3 a_2}{|a|^2} & \frac{a_3^2}{|a|^2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

EXAMPLE 21.14. The projection $Px = x - \frac{x \cdot n}{|n|^2} n$ onto a plane $n \cdot x = 0$ through the origin takes the matrix form

$$Px = \begin{pmatrix} 1 - \frac{n_1^2}{|n|^2} & -\frac{n_1 n_2}{|n|^2} & -\frac{n_1 n_3}{|n|^2} \\ -\frac{n_2 n_1}{|n|^2} & 1 - \frac{n_2^2}{|n|^2} & -\frac{n_2 n_3}{|n|^2} \\ -\frac{n_3 n_1}{|n|^2} & -\frac{n_3 n_2}{|n|^2} & 1 - \frac{n_3^2}{|n|^2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

EXAMPLE 21.15. The mirror image of a point x with respect to a plane through the origin given by $(2P - I)x$, where Px is the projection of x onto the plane, takes the matrix form

$$(2P - I)x = \begin{pmatrix} 2\frac{a_1^2}{|a|^2} - 1 & 2\frac{a_1 a_2}{|a|^2} & 2\frac{a_1 a_3}{|a|^2} \\ 2\frac{a_2 a_1}{|a|^2} & 2\frac{a_2^2}{|a|^2} - 1 & 2\frac{a_2 a_3}{|a|^2} \\ 2\frac{a_3 a_1}{|a|^2} & 2\frac{a_3 a_2}{|a|^2} & 2\frac{a_3^2}{|a|^2} - 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

21.31 The scalar product is invariant under orthogonal transformations

Let Q be the matrix $\{q_1, q_2, q_2\}$ formed by taking the columns to be the basis vectors q_j. We assume that Q is orthogonal, which is the same as assuming that $\{q_1, q_2, q_2\}$ is an orthogonal basis, that is the q_j are pairwise orthogonal and have length 1. The coordinates $\hat{x}$ of a vector x in the standard basis with respect to the basis $\{q_1, q_2, q_2\}$ are given by $\hat{x} = Q^{-1}x = Q^\top x$. We shall now prove that if $\hat{y} = Q^\top y$, then

$$\hat{x} \cdot \hat{y} = x \cdot y,$$

which states that the scalar product is invariant under orthogonal coordinate changes. We compute

$$\hat{x} \cdot \hat{y} = (Q^\top x) \cdot (Q^\top y) = x \cdot (Q^\top)^\top Q^\top y = x \cdot y,$$

where we used that for any 3×3 matrix $A = (a_{ij})$ and $x, y \in \mathbb{R}^3$

$$\begin{aligned}(Ax) \cdot y &= \sum_{i=1}^{3}(\sum_{j=1}^{3} a_{ij}x_j)y_i = \sum_{j=1}^{3}(\sum_{i=1}^{3} a_{ij}y_i)x_j \\ &= (A^\top y) \cdot x = x \cdot (A^\top y),\end{aligned} \tag{21.45}$$

with $A = Q^\top$, and the facts that $(Q^\top)^\top = Q$ and $QQ^\top = I$.

We can now complete the argument about the geometric interpretation of the scalar product from the beginning of this chapter. Given two non-parallel vectors a and b, we may assume by an orthogonal coordinate transformation that a and b belong to the $x_1 - x_2$-plane and the geometric interpretation from Chapter *Analytic geometry in* $\mathbb{R}^2$ carries over.

21.32 Looking ahead to functions $f : \mathbb{R}^3 \to \mathbb{R}^3$

We have met linear transformations $f : \mathbb{R}^3 \to \mathbb{R}^3$ of the form $f(x) = Ax$, where A is a 3×3 matrix. Below we shall meet more general (non-linear) transformations $f : \mathbb{R}^3 \to \mathbb{R}^3$ that assign a vector $f(x) = (f_1(x), f_2(x), f_3(x)) \in \mathbb{R}^3$ to each $x = (x_1, x_2, x_3) \in \mathbb{R}^3$. For example,

$$f(x) = f(x_1, x_2, x_3) = (x_2x_3, x_1^2 + x_3, x_3^4 + 5)$$

with $f_1(x) = x_2x_3$, $f_2(x) = x_1^2+x_3$, $f_3(x) = x_3^4+5$. We shall see that we may naturally extend the concepts of Lipschitz continuity and differentiability for functions $f : \mathbb{R} \to \mathbb{R}$ to functions $f : \mathbb{R}^3 \to \mathbb{R}^3$. For example, we say that $f : \mathbb{R}^3 \to \mathbb{R}^3$ is Lipschitz continuous on $\mathbb{R}^3$ if there is a constant L_f such that

$$|f(x) - f(y)| \leq L_f|x - y| \quad \text{for all } x, y \in \mathbb{R}^3.$$

21.33 Unity of Body and Soul

The history of the study the relation of Body and Soul since Descartes largely concerns attempts to escape the Cartesian dualism or interactionism. In his metaphysical masterpiece *De Ethica*, Spinoza (1632-1677) put forward the *double-aspect theory* stating that the mental and the physical aspects are simply different aspects of *one and the same single substance*, which for Spinoza was God, the universal essence or nature of everything that exists.

In *Systéme nouveau de la nature* and *Eclaircissement du nouveau sisteme* from 1695-96, Leibniz presented his famous *psychophysical parallelism* with the view that Soul and Body exist in a *Pre-established Harmony*: every mental event is correlated with a physical event in such a way that when one occurs, so too does the other. More precisely, Leibniz states in *Monadology* from 1714: "The Soul follows its own laws and the Body also follows its own; and they agree in virtue of the harmony pre-established between all substances, since they are all represenations of a single universe... Souls act according to the laws of *final causes*, through appetitions, ends and means. Bodies act according to the laws of *efficient causes* or of motions. And these two kingdoms, that of efficient causes and that of final causes, are in harmony with each other...According to this system, Bodies act as if there were no Souls, and Souls act as if there were no Bodies, and both act as if they influenced the other".

Also, in a letter from 1697, Leibniz writes:"I believe that everything really happens mechanically in nature; and can be explained by efficient causes, but that, at the same time, everything also takes place morally, so to speak, and can be explained by final causes. These two kingdoms, the moral one of minds and souls and the mechanical one of bodies, penetrate each other and are in perfect accord through the agency of the Author of things, who is at the same time the first efficient cause and the last end".

During the 19th century the view that the brain serves as the organ of mind/soul developed and different theories for mind/brain relations were presented including immaterialism. materialism, dual-aspect monism, and mind-stuff theory, all of them seeking to circumvent the Cartesian dualistic interactionism.

Immaterialism, best represented by George Berkeley (1685-1753) in his *A Treatise concerning the Principles of Human Knowledge* from 1710, denies even the possibility of mindless material substance. For something to exist for Berkeley, it must either be perceived or be the active mind doing the perceiving. From this perspective, there is no mind/body distinction because what we think of as body is merely the perception of mind. Immaterialism was to resurface in the later 19th century in the guise of mind-stuff theory.

Materialism, which dates to antiquity, holds that matter is fundamental. Whatever else may exist, if it exists, it depends on matter. This view was presented by la Mettrie (1709-1751) in *Histoire naturelle de l'ame* and *L'homme machine* with an extension of Descartes' automata concept from animals to man. In particular, L'homme machine introduced the critical notion that conscious and voluntary processes are only distinguished from involuntary and instinctual activities by the relative complexity of their mechanical substrate, and presented the idea of a *living machine* as a purposive, autonomous, and dynamic system.

Cabanis (1757-1808) continued in the same spirit with *Rapports du physique et du moral de l'homme* from 1802, arguing that "to have an accurate idea of the operations from which thought results, it is necessary to consider the brain as a special organ designed especially to produce it, as the stomach and the intestines are designed to operate the digestion, (and) the liver to filter bile..."

Thomas Henry Huxley (1825-1895) presented related ideas in *On the hypothesis that animals are automata, and its history* from 1874, suggesting that states of consciousness are merely the effect of molecular changes in brain substance that has attained a prerequisite degree of organization. Animals, therefore, are ”conscious automata.”

Dual-aspect monism was the brain child of George Henry Lewes (1817-1878), biologist, philosopher, and psychologist, articulated in *The Physical Basis of Mind*: there is only one kind of ”stuff” and that mind and body differ only in the arrangement of that stuff or in the perspective from which it is apprehended.

Mind-stuff theory was further developed by Morton Prince (1854-1929) in *The Nature of Mind and Human Automatism* from 1885, stating that higher properties of mind, such as judgment, reasoning, volition, or the continuous flow of consciousness, are compounded from mental elements (pieces of mind-stuff) that do not in themselves manifest these higher properties. Even the most basic material elements possess a small piece of mind-stuff such that when these elements are combined, mind-stuff is similarly combined.

In *The Principles of Psychology* Herbert Spencer (1820-1903) stressed three basic evolutionary principles of *adaptation*, *continuity* and *development*: ”All those activities, bodily and mental, which constitute our ordinary idea of life, consist simply of the continuous adjustment of internal relations to external relations. Mental phenomena are defined as adaptations, incidents of the correspondence between the organism and its environment”. There is thus no justification for drawing a sharp distinction between mind and body.

We continue the story of Body and Soul in Chapter *Analytic geometry in* $\mathbb{R}^n$ below with basic questions of Artificial Intelligence AI such as: Can a computer think?

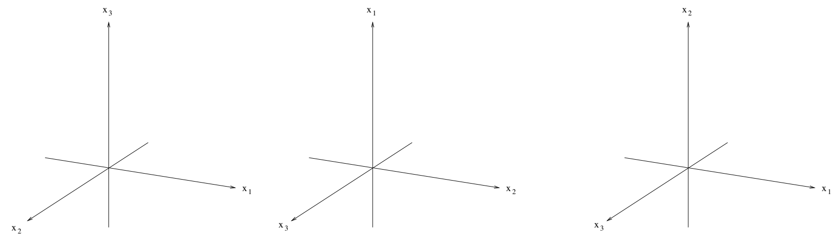
Chapter 21 Problems

21.1. Show that the norm $|a|$ of the vector $a = (a_1, a_2, a_3)$ is equal to the distance from the origin $0 = (0, 0, 0)$ to the point (a_1, a_2, a_3). Hint: apply Pythagoras Theorem twice.

21.2. Which of the following coordinate systems are righthanded?

21.3. Indicate the direction of $a \times b$ and $b \times a$ in Fig. 21.1 if b points in the direction of the x_1-axis. Consider also the same question in Fig. 21.2.

21.4. Given $a = (1, 2, 3)$ and $b = (1, 3, 1)$, compute $a \times b$.



21.5. Compute the volume of the parallelepiped spanned by the three vectors $(1, 0, 0)$, $(1, 1, 1)$ and $(-1, -1, 1)$.

21.6. What is the area of the triangle formed by the three points: $(1, 1, 0)$, $(2, 3, -1)$ and $(0, 5, 1)$?

21.7. Given $b = (1, 3, 1)$ and $a = (1, 1, 1)$, compute a) the angle between a and b, b) the projection of b onto a, c) a unit vector orthogonal to both a and b.

21.8. Consider a plane passing through the origin with normal $n = (1, 1, 1)$ and a vector $a = (1, 2, 3)$. Which point p in the plane has the shortest distance to a?

21.9. Is it true or not that for any 3×3 matrices A, B, and C and number λ (a) $A + B = B + A$, (b) $(A + B) + C = A + (B + C)$, (c) $\lambda(A + B) = \lambda A + \lambda B$?

21.10. Prove that for 3×3 matrices A, B and C: $(AB)C = A(BC)$. Hint: Use that $D = (AB)C$ has the elements $d_{ij} = \sum_{k=1}^{3}(\sum_{l=1}^{3} a_{il}b_{lk})c_{kj}$, and do the summation in a different order.

21.11. Give examples of 3×3-matrices A and B such that $AB \neq BA$. Is it difficult to find such examples, that is, is it exceptional or "normal" that $AB \neq BA$.

21.12. Prove Theorem 21.5.

21.13. Write down the three matrices corresponding to rotations around the x_1, x_2 and x_3 axis.

21.14. Find the matrix corresponding to a rotation by the angle θ around a given vector b in $\mathbb{R}^3$.

21.15. Give the matrix corresponding to be mirroring a vector through the $x_1 - x_2$-plane.

21.16. Consider a linear transformation that maps two points p_1 and p_2 in $\mathbb{R}^3$ into the points $\hat{p}_1$, $\hat{p}_2$, respectively. Show that all points lying on a straight line between p_1, p_2 will be transformed onto a straight line between $\hat{p}_1$ and $\hat{p}_2$.

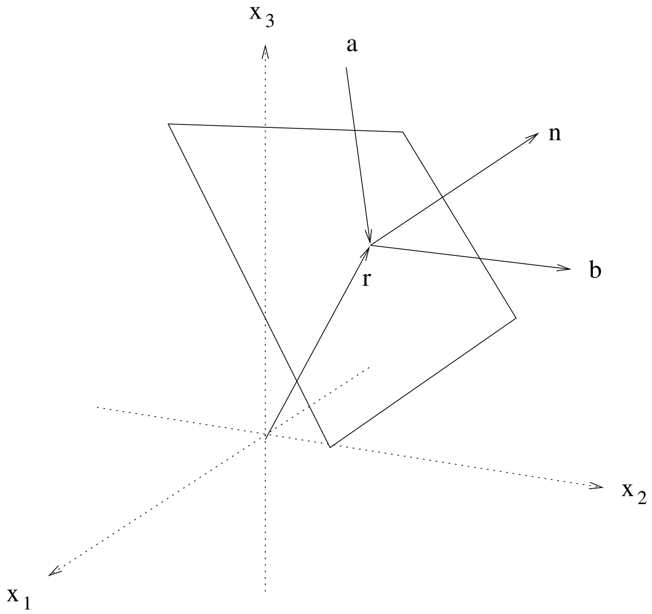
21.17. Consider two straight lines in $\mathbb{R}^3$ given by: $a + \lambda b$ and $c + \mu d$ where $a, b, c, d \in \mathbb{R}^3$, $\lambda, \mu \in \mathbb{R}$. What is the shortest distance between the two lines?

21.18. Compute the intersection of the two lines given by: $(1, 1, 0) + \lambda(1, 2, -3)$ and $(2, 0, -3) + \mu(1, 1, -3)$. Is it a rule or an exception that such an intersection can be found?

21.19. Compute the intersection between two planes passing through the origin with normals $n_1 = (1,1,1)$, $n_2 = (2,3,1)$. Compute the intersection of these two planes and the $x_1 - x_2$ plane.

21.20. Prove that (21.42) implies that the inverse of A^{-1} exists.

21.21. Consider a plane through a point r with normal n. Determine the reflection in the plane at r of a light ray entering in a direction parallel to a given vector a.



21.22. Show that the distance between a point b and its projection onto a plane $n \cdot x = d$ is equal to the shortest distance between b and any point in the plane. Give both a geometric proof based on Pythagoras' theorem, and an analytical proof. Hint: For x in the plane write $|b-x|^2 = |b - Pb + (Pb-x)|^2 = (b - Pb + (Pb-x), b - Pb + (Pb-x))$ and expand using that $(b - Pb, Pb - x) = 0$).

21.23. Express (21.31) i matrix form.

21.24. Complete the proof of the claim that (21.30) is uniquely solvable.

> ...Show that the meridian, hce che ech, interecting at royde angles the parilegs of a given obtuse one biscuts both the arcs and are in curveachord behind. Brickbaths. The family umbroglia. A Tullagrove pole to the Height of County Fearmanagh has a septain inclinaison and the graphplot for all the functions in Lower County Monachan, whereat samethings is rivisble by nighthim, may be involted into the

zeroic couplet, palls pell inhis heventh glike noughty times ∞, find, if you rae not literally cooefficient, how minney combinaisies and permutandies can be played on the international surd! pthwndxrclzp!, hids cubid rute being extructed, taking anan illitterates, ifif at a tom. Answers, (for teasers only). Ten, twent, thirt, see, ex and three icky totchty ones. from solation to solution. Imagine the twelve deaearended dumbbawls of the whowl abovebeugled to be the contonuation through regeneration of the urutteration of the word in pregross. It follows that, if the two antesedents be bissyclitties and the three comeseekwenchers trundletrikes, then, Aysha Lalipat behidden on the footplate, Big Whiggler restant upsittuponable, N$\mathbb{CR}$ presents to us (tandem year a lasted length) an ottomanic turquoindaco of pictorial shine by pictorial shimmer so long as, gad of gidday, pictorial summer, viridorefulvid, lits asheen, but (lenz alack lends a lot), if this habby cyclic redor be outraciously enviolated by a mierelin roundtablerunning, like knuts in maze, the zitas runnind hare and dart with the yeggs in their muddle, like a seven of wingless arrows, hodgpadge, thump, kick and hurry, all boy more missis blong him he race quickfeller all some hogglepiggle longer house blong him, while the catched and dodged exarx seems himmulteemiously to beem (he wins her hend! he falls to tail!, the erset ladest mandand (uhu and uhud) the loose frace on erroroots, twalegged poneys and threehandled dorkeys, (madahoy, morahoy lugahoy, jogahoyaway) MPM brings us a rainborne pamtomomiom, aqualavant to (Cat my dogs, if I baint dingbushed leik everything!) kaksitoista, volts yksitoista volts kymmenen volts yhdeksan volts kahdeksan volts seitseman volts kuusi volys viisi volts nejla volts kolme volts volts kaksi volts yksi volts, allahthallacamellated, caraven series to the finish of helvé fractures. In outher wards, one from five, one from fives two, two to fives ones millamills with mill and a half a mill a twos fives five of bullyclavers. For a surview over all the factionables see Iris in the Evenine's World. Binomeans to be comprehended. Inexcessible as thy by god ways. The aximones. Ande their prostalutes. For his neuralgiabrown. Equal to = aosch. (Finnegans Wake, James Joyce)

22
Complex numbers

The imaginary number is a fine and wonderful recourse of the divine spirit, almost an amphibian between being and not being. (Leibniz)

The composition of vast books is a laborious and impoverishing extravagance. To go on for five hundred pages developing an idea whose perfect oral exposition is possible in a few minutes! A better course of procedure is to pretend that these books already exist, and then to offer a resume, a commentary...More reasonable, more inept, more indolent, I have preferred to write notes upon imaginary books. (Borges, 1941)

22.1 Introduction

In this chapter, we introduce the set of *complex numbers* $\mathbb{C}$. A complex number, typically denoted by z, is an ordered pair $z = (x, y)$ of real numbers x and y, where x represents the *real part* of z and y the *imaginary part* of z. We may thus identify $\mathbb{C}$ with $\mathbb{R}^2$ and we often refer to $\mathbb{C}$ as the *complex plane*. We further identify the set of complex numbers with zero imaginary part with the set of real numbers and write $(x, 0) = x$, viewing the real line $\mathbb{R}$ as the x-axis in the complex plane $\mathbb{C}$. We may thus view $\mathbb{C}$ as an extension of $\mathbb{R}$. Similarly, we identify the set of complex numbers with zero real part with the y-axis, which we also refer to as the set of *purely imaginary* numbers. The complex number $(0, 1)$ is given a special name $i = (0, 1)$, and we refer to i as the *imaginary unit*.

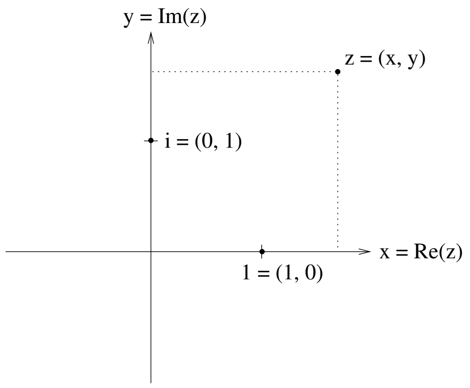


FIGURE 22.1. The complex plane $\mathbb{C} = \mathbb{R}^2$

The operation of addition in $\mathbb{C}$ coincides with the operation of vector addition in $\mathbb{R}^2$. The new aspect of $\mathbb{C}$ is the operation of multiplication of complex numbers, which differs from scalar and vector multiplication in $\mathbb{R}^2$.

The motivation to introduce complex numbers comes from considering for example the polynomial equation $x^2 = -1$, which has no root if x is restricted to be a real number. There is no real number x such that $x^2 = -1$ since $x^2 \geq 0$ for $x \in \mathbb{R}$. We shall see that if we allow x to be a complex number, the equation $x^2 = -1$ becomes solvable and the two roots are $x = \pm i$. More generally, the Fundamental Theorem of Algebra states that any polynomial equation with real or complex coefficients has a root in the set of complex numbers. In fact, it follows that a polynomial equation of degree n has exactly n roots.

Introducing the complex numbers finishes the extension process from natural numbers over integers and rational numbers to real numbers, where in each case a new class of polynomial equations could be solved. Further extensions beyond complex numbers to for example *quarternions* consisting of quadruples of real numbers were made in the 19th century by Hamilton, but the initial enthusiasm over these constructs faded since no fully convincing applications were found. The complex numbers, on the other hand, have turned out to be very useful.

22.2 Addition and multiplication of complex numbers

We define the *sum* $(a, b) + (c, d)$ of two complex numbers (a, b) and (c, d), obtained through the operation of *addition* denoted by $+$, as follows:

$$(a, b) + (c, d) = (a + c, b + d), \tag{22.1}$$

that is we add the real parts and imaginary parts separately. We see that addition of two complex numbers corresponds to vector addition of the corresponding ordered pairs or vectors in $\mathbb{R}^2$. Of course, we define subtraction similarly: $(a,b)-(c,d)=(a-c,b-d)$.

We define the *product* $(a,b)(c,d)$ of two complex numbers (a,b) and (c,d), obtained through the operation of *multiplication*, as follows:

$$(a,b)(c,d)=(ac-bd,ad+bc). \tag{22.2}$$

We can readily check using rules for operating with real numbers that the operations of addition and multiplication of complex numbers obey the commutative, associative and distributive rules valid for real numbers.

If $z=(x,y)$ is a complex number, we can write

$$z=(x,y)=(x,0)+(0,y)=(x,0)+(0,1)(y,0)=x+iy, \tag{22.3}$$

referring to the identification of complex numbers of the form $(x,0)$ with x, (and similarly $(y,0)$ with y of course) and the notation $i=(0,1)$ introduced above. We refer to x is the *real part of* z and y as the *imaginary part of* z, writing $x=\text{Re } z$ and $y=\text{Im } z$, that is

$$z=\text{Re } z+i\,\text{Im } z=(\text{Re } z,\text{Im } z). \tag{22.4}$$

We note in particular that

$$i^2=i\,i=(0,1)(0,1)=(-1,0)=-(1,0)=-1, \tag{22.5}$$

and thus $z=i$ solves the equation $z^2+1=0$. Similarly, $(-i)^2=-1$, and thus the equation $z^2+1=0$ has the two roots $z=\pm i$.

The rule (22.2) for multiplication of two complex numbers (a,b) and (c,d), can be retrieved using that $i^2=-1$ (and taking the distributive law for granted):

$$(a,b)(c,d)=(a+ib)(c+id)=ac+i^2bd+i(ad+bc)=(ac-bd,ad+bc).$$

We define the *modulus* or absolute value $|z|$ of a complex number $z=(x,y)=x+iy$, by

$$|z|=(x^2+y^2)^{1/2}, \tag{22.6}$$

that is, $|z|$ is simply the length or norm of the corresponding vector $(x,y)\in\mathbb{R}^2$. We note that if $z=x+iy$, then in particular

$$|x|=|\text{Re } z|\le|z|, \qquad |y|=|\text{Im } z|\le|z|. \tag{22.7}$$

22.3 The triangle inequality for complex numbers

If z_1 and z_2 are two complex numbers, then

$$|z_1+z_2|\le|z_1|+|z_2|. \tag{22.8}$$

This is the *triangle inequality for complex numbers*, which follows directly from the triangle inequality in $\mathbb{R}^2$.

22.4 Open domains in $\mathbb{C}$

We extend the notion of an open domain in $\mathbb{R}^2$ to $\mathbb{C}$ in the natural way. We say that a domain Ω in $\mathbb{C}$ is *open* if the corresponding domain in $\mathbb{R}^2$ is open, that is for each $z_0 \in \Omega$ there is a positive number r_0 such that the complex numbers z with $|z - z_0| < r$ also belong to Ω. For example, the set $\Omega = \{z \in \mathbb{C} : |z| < 1\}$, is open.

22.5 Polar representation of complex numbers

Using polar coordinates in $\mathbb{R}^2$, we can express a complex number as follows

$$z = (x, y) = r(\cos(\theta), \sin(\theta)) = r(\cos(\theta) + i\sin(\theta)), \qquad (22.9)$$

where $r = |z|$ is the modulus of z and $\theta = \arg z$ is the *argument* of z, and we also used (22.3). We usually assume that $\theta \in [0, 2\pi)$, but by periodicity we may replace θ by $\theta + 2\pi n$ with $n = \pm 1, \pm 2, ...,$. Choosing $\theta \in [0, 2\pi)$, we obtain the *principal argument* of z, which we denote by Arg z.

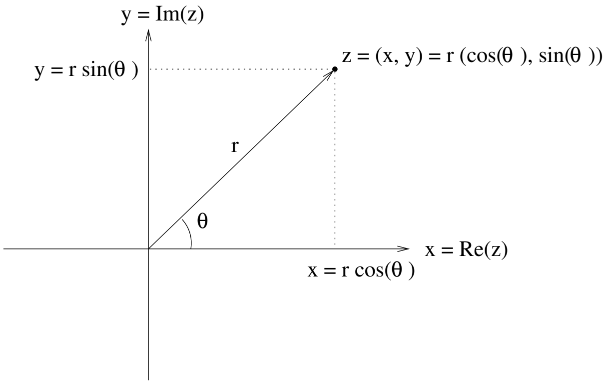


FIGURE 22.2. Polar representation of a complex number

EXAMPLE 22.1. The polar representation of the complex number $z = (1, \sqrt{3}) = 1 + i\sqrt{3}$ is $z = 2(\cos(\frac{\pi}{3}), \sin(\frac{\pi}{3}))$, or $z = 2(\cos(60°), \sin(60°))$.

22.6 Geometrical interpretation of multiplication of complex numbers

To find the operation on vectors in $\mathbb{R}^2$ corresponding to multiplication of complex numbers, it is convenient to use polar coordinates,

$$z = (x, y) = r(\cos(\theta), \sin(\theta)),$$

where $r = |z|$ and $\theta = \text{Arg } z$. Letting $\zeta = (\xi, \eta) = \rho(\cos(\varphi), \sin(\varphi))$ be another complex number expressed using polar coordinates, the basic trigonometric formulas from the Chapter Pythagoras and Euclid imply

$$\begin{aligned} z\zeta &= r(\cos(\theta), \sin(\theta))\, \rho(\cos(\varphi), \sin(\varphi)) \\ &= r\rho(\cos(\theta)\cos(\varphi) - \sin(\theta)\sin(\varphi), \cos(\theta)\sin(\varphi) + \sin(\theta)\cos(\varphi)) \\ &= r\rho(\cos(\theta + \varphi), \sin(\theta + \varphi)). \end{aligned}$$

We conclude that multiplying $z = r(\cos(\theta), \sin(\theta))$ by $\zeta = \rho(\cos(\varphi), \sin(\varphi))$ corresponds to rotating the vector z the angle $\varphi = \text{Arg } \zeta$, and changing its modulus by the factor $\rho = |\zeta|$. In other words, we have

$$\arg z\zeta = \text{Arg } z + \text{Arg } \zeta, \quad |z\zeta| = |z||\zeta|. \tag{22.10}$$

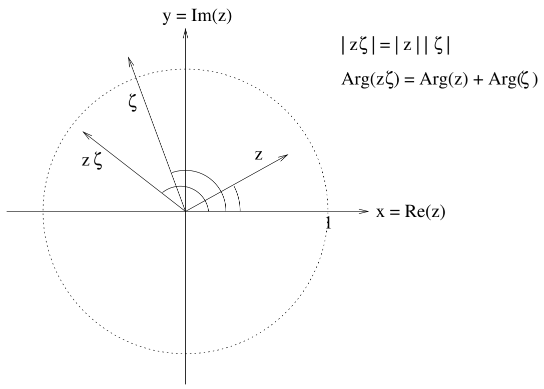


FIGURE 22.3. Geometrical interpretation of multiplication of a complex numbers

EXAMPLE 22.2. Multiplication by i corresponds to rotation counterclockwise $\frac{\pi}{2}$, or 90°.

22.7 Complex conjugation

If $z = x + iy$ is a complex number with x and y real, we define the complex conjugate $\bar{z}$ of z as

$$\bar{z} = x - iy.$$

We see that z is real if and only if $\bar{z} = z$ and that z is *purely imaginary*, that is Re z $= 0$, if and only $z = -\bar{z}$.

Identifying $\mathbb{C}$ with $\mathbb{R}^2$, we see that complex conjugation corresponds to reflection in the real axis. We also note the following relations, easily verified,

$$|z|^2 = z\bar{z}, \quad \text{Re } z = \frac{1}{2}(z + \bar{z}), \quad \text{Im } z = \frac{1}{2i}(z - \bar{z}). \tag{22.11}$$

22.8 Division of complex numbers

We extend the operation of division (denoted by /) of real numbers to division of complex numbers by defining for $w, u \in \mathbb{C}$ with $u \neq 0$,

$$z = w/u = \frac{w}{u} \quad \text{if and only if } uz = w.$$

To compute w/u for given $w, u \in \mathbb{C}$ with $u \neq 0$, we proceed as follows:

$$w/u = \frac{w}{u} = \frac{w\bar{u}}{u\bar{u}} = \frac{w\bar{u}}{|u|^2}.$$

EXAMPLE 22.3. We have

$$\frac{1+i}{2+i} = \frac{(1+i)(2-i)}{5} = \frac{3}{5} + i\frac{1}{5}.$$

Note that we consider complex numbers as *scalars* although they have a lot in common with vectors in $\mathbb{R}^2$. The main reason for this is that ....

22.9 The Fundamental Theorem of Algebra

Consider a polynomial equation $p(z) = 0$, where $p(z) = a_0 + a_1 z + ... + a_n z^n$ is a polynomial in z of degree n with complex coefficients $a_0, ..., a_n$. The Fundamental Theorem of Algebra states that the equation $p(z)$ has at least one complex root z_1 satisfying $p(z_1) = 0$. By the factorization algorithm, it follows that $p(z)$ can be factored into

$$p(z) = (z - z_1)p_1(z),$$

where $p_1(z)$ is a polynomial of degree at most $n-1$. Indeed, the factorization algorithm from the Chapter Combinations of functions (Section 11.4) shows that

$$p(z) = (z - z_1)p_1(z) + c,$$

where c is a constant. Setting $z = z_1$, it follows that $c = 0$. Repeating the argument, we find that $p(z)$ can be factored into

$$p(z) = c(z - z_1)....(z - z_n),$$

where $z_1, .., z_n$ are the (complex valued in general) roots of $p(z) = 0$.

22.10 Roots

Consider the equation in $w \in \mathbb{C}$

$$w^n = z,$$

where $n = 1, 2, ..$ is a natural number and $z \in \mathbb{C}$ is given. Using polar coordinates with $z = |z|(\cos(\theta), \sin(\theta)) \in \mathbb{C}$ and $w = |w|(\cos(\varphi), \sin(\varphi)) \in \mathbb{C}$, the equation $w^n = z$ takes the form

$$|w|^n(\cos(n\varphi), \sin(n\varphi)) = |z|(\cos(\theta), \sin(\theta))$$

from which it follows that

$$|w| = |z|^{\frac{1}{n}}, \qquad \varphi = \frac{\theta}{n} + 2\pi\frac{k}{n},$$

where $k = 0, ..., n-1$. We conclude that the equation $w^n = z$ has n distinct roots on the circle $|w| = |z|^{\frac{1}{n}}$. In particular, the equation $w^2 = -1$ has the two roots $w = \pm i$. The n roots of the equation $w^n = 1$ are called the n *roots of unity.*

22.11 Solving a quadratic equation $w^2 + 2bw + c = 0$

Consider the quadratic equation for $w \in \mathbb{C}$,

$$w^2 + 2bw + c = 0,$$

where $b, c \in \mathbb{C}$. Completing the square, we get

$$(w + b)^2 = b^2 - c.$$

If $b^2 - c \geq 0$ then

$$w = -b \pm \sqrt{b^2 - c},$$

while if $b^2 - c < 0$ then

$$w = -b \pm i\sqrt{c - b^2}.$$

22.12 Gösta Mittag-Leffler

The Swedish mentor of Sonya Kovalevskaya was Gösta Mittag-Leffler (1846-1927), famous Swedish mathematician and founder of the prestigous journal Acta Mathematica, see Fig. 22.4. The huge mansion of Mittag-Leffler, beautifully situated in in Djursholm, just outside Stockholm, with an impressive library, now houses Institut Mittag-Leffler bringing mathematicians from all over the world together for work-shops on different themes of mathematics and its applications. Mittag-Leffler made important contributions to the theory of functions of a complex variable, see Chapter *Analytic functions* below.

FIGURE 22.4. Gösta Mittag-Leffler, Swedish mathematician and founder of Acta Mathematica: "The mathematician's best work is art, a high perfect art, as daring as the most secret dreams of imagination, clear and limpid. Mathematical genius and artistic genius touch one another".

Chapter 22 Problems

22.1. Show that (a) $\frac{1}{i} = -i$, (b) $i^4 = 1$.

22.2. Find (a) Re $\frac{1}{1+i}$, (b) Im $\frac{3+4i}{7-i}$, (c) Im $\frac{z}{\bar{z}}$.

22.3. Let $z_1 = 4 - 5i$ and $z_2 = 2 + 3i$. Find in the form $z = x + iy$ (a) $z_1 z_2$, (b) $\frac{z_1}{z_2}$, (c) $\frac{z_1}{z_1 + z_2}$.

22.4. Show that the set of complex numbers z satisfying an equation of the form $|z - z_0| = r$, where $z_0 \in \mathbb{C}$ is given and $r > 0$, is a circle in the complex plane with center z_0 and radius r.

22.5. Represent in polar form (a) $1 + i$, (b) $\frac{1+i}{1-i}$, (c) $\frac{2+3i}{5+4i}$.

22.6. Solve the equations (a) $z^2 = i$, (b) $z^8 = 1$, (c) $z^2 + z + 1 = -i$, (d) $z^4 - 3(1 + 2i)z^2 + 6i = 0$.

22.7. Determine the sets in the complex plane represented by (a) $|\frac{z+i}{z-i}| = 1$, (b) $\mathrm{Im}\ z^2 = 2$, (c) $|\mathrm{Arg}\ z| \leq \frac{\pi}{4}$.

22.8. Express z/w in polar coordinates in terms of the polar coordinates of z and w.

22.9. Describe in geometrical terms the mappings $f : \mathbb{C} \to \mathbb{C}$ given by (a) $f(z) = az + b$, with $a, b \in \mathbb{C}$, (b) $f(z) = z^2$, (c) $f(z) = z^{\frac{1}{2}}$.

23
The derivative

I'll teach you differences. (Shakespeare: King Lear)

An object with zero velocity will not change position. (Einstein)

... and therefore I offer this work as the mathematical principles of philosophy, for the whole burden in philosophy seems to consist in this: from the phenomena of motions to investigate the forces of nature, and then from these forces to demonstrate the other phenomena ...(Galileo)

23.1 Rates of change

Life is change. The newborn changes every day and acquires new skills, the teen-ager develops into an adult in a couple of years, the middle-aged wants to see the family, the house and career expand every year. Only the retired wants to stop the world and play golf for ever, but realizes that this is impossible and understands that there is an end, after which there is no change at all any more.

When something changes, we may speak of the *total change* and we may speak of the *change per unit* or the *rate of change*. If our salary increases, we expect an increase in tax and we may speak of the total change in tax (for one year). We may also speak of the change in tax per extra dollar we earn, which is a rate of change of tax commonly referred to as *marginal income tax*. The marginal income tax usually changes with our total income, so that we pay a higher marginal tax if we have a higher income. If our total income is 10 000 dollars, then we may have to pay 30 cents tax out of an extra dollar we earn, and if our total income is 50 000 dollars, we may have to pay 50 cents tax out of an extra dollar. The marginal tax, or rate of change of tax, in this example is 0.3 if our income is 10 000 dollars and 0.5 if our income is 50 000 dollars.

Business people speak of *marginal cost* of a certain item, which is the increase in total cost if we buy one more item, that is the cost increase per item or rate of change of total cost. Normally the marginal cost depends on the the total amount and in fact normally the marginal cost decreases with the total amount of items we buy. The marginal cost of producing some item also varies with the total amount produced. At a certain production level, the cost of producing one more item may be very small, while if we have to build a whole new factory to produce that single additional item, the marginal cost would be very large. Thus the marginal cost in production may vary with the total production.

The concept of a function $f : D(f) \to R(f)$ is also intimately connected to change. For each $x \in D(f)$ there is a $f(x) \in R(f)$, and usually $f(x)$ changes with x. If $f(x)$ is the same for all x, then the function $f(x)$ is a constant function, which is easy to grasp and does not require much further study. If $f(x)$ does vary with x, then it is natural to seek ways of describing qualitatively and quantitatively how $f(x)$ varies with x. The rate of change enters again if we seek to describe how $f(x)$ changes per unit of x.

The *derivative* of a function $f(x)$ with respect to x measures the rate of change of $f(x)$ as x varies. The derivative of our tax with respect to income is the marginal tax. The derivative of the total production cost with respect to total production is the marginal cost.

The basic modeling tool in Calculus is the derivative. Indeed, the start of the modern scientific age coincides with the invention of the concept of derivative. The derivative is a measure of rate of change.

In this chapter, we introduce the wonderful mathematical concept of the derivative, figure out some of its properties, and start to use derivatives in mathematical modeling.

23.2 Paying taxes

We return to the above example of describing our income tax as a function of income. Suppose we let x denote our total income next year and let $f(x)$ be the corresponding total income tax we would have to pay. The function $f(x)$ describes how our total income tax changes with our income x. For each given income x, there is a corresponding income tax $f(x)$ to pay. We plot a possible function $f(x)$ in the following figure:

The function $f(x)$ in this example is piecewise linear and Lipschitz continuous. The slope of $f(x)$ is zero up to the total income 5, the slope is 0.2

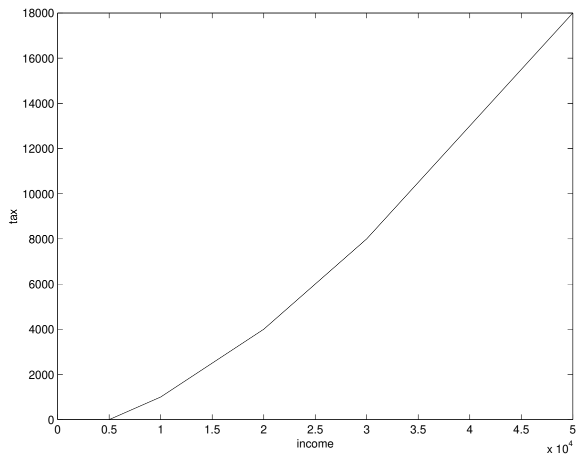


FIGURE 23.1. Income tax $f(x)$ varying with income x.

in the interval $[5\,000, 10\,000]$, 0.3 in $[10\,000, 20\,000]$, 0.4 in $[20\,000, 30\,000]$ and 0.5 in $[30\,000, \infty)$.

For a given $\bar{x}$, the slope of the straight line representing $f(x)$ close to $\bar{x}$, is the marginal tax. We denote the slope of $f(x)$ at $\bar{x}$ by $m(\bar{x})$. We see that the slope $m(\bar{x})$ varies with $\bar{x}$. For example, $m(\bar{x}) = 0.3$ for $\bar{x} \in (10\,000, 20\,000)$. If we add one extra dollar at the income $\bar{x} \in (10\,000, 20\,000)$, then our income tax will increase by 0.3 dollars.

The marginal tax is the same as the slope of the straight line representing the income tax $f(x)$ as a function of income x. Thus the marginal tax is $m(\bar{x})$ at the income $\bar{x}$. The marginal income tax is zero up to the total income 5 000, the marginal tax is 0.2 in the income bracket $[5\,000, 10\,000]$, 0.3 in the bracket $[10\,000, 20\,000]$, 0.4 in the bracket $[20\,000, 30\,000]$ and 0.5 for incomes in the bracket $[30\,000, \infty)$.

We can describe how $f(x)$ varies in each income tax bracket through the following formula

$$
\begin{aligned}
&f(x) = 0 \quad \text{for } x \in [0, 5\,000]\\
&f(x) = 0.2(x - 5\,000) \quad \text{for } x \in [5\,000, 10\,000]\\
&f(x) = f(10\,000) + 0.3(x - 10\,000) \quad \text{for } x \in [10\,000, 20\,000]\\
&f(x) = f(20\,000) + 0.4(x - 20\,000) \quad \text{for } x \in [20\,000, 30\,000]\\
&f(x) = f(30\,000) + 0.5(x - 30\,000) \quad \text{for } x \in [30\,000, \infty)
\end{aligned}
$$

We can condense these formulas into

$$f(x) = f(\bar{x}) + m(\bar{x})(x - \bar{x}), \tag{23.1}$$

where $\bar{x}$ represents a given income with corresponding tax $f(\bar{x})$, and we are interested in the tax $f(x)$ for an income x in some interval containing

$\bar{x}$. For example, the formula

$$f(x) = f(15\,000) + m(15\,000)(x - 15\,000) \quad \text{for } x \in [10\,000, 20\,000]$$

where $m(15\,000) = 0.3$ is the marginal tax, describes how the tax varies with the income x around the income $\bar{x} = 15\,000$, see Fig. 23.2.

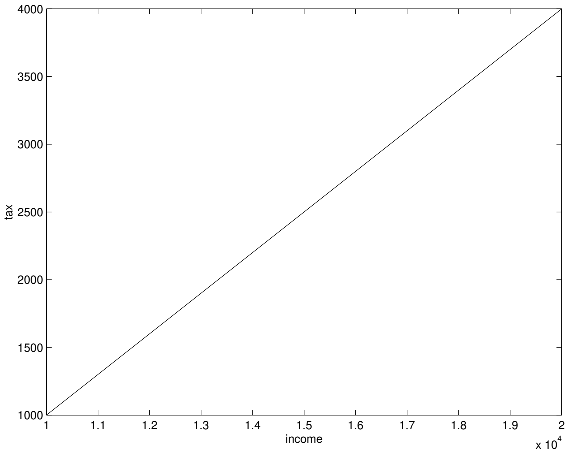


FIGURE 23.2. Income tax $f(x)$ for income x in the interval $[10\,000, 20\,000]$.

The derivative of the function $f(x) = f(\bar{x}) + m(\bar{x})(x - \bar{x})$ for $x = \bar{x}$, is the marginal tax $m(\bar{x})$. The formula $f(x) = f(\bar{x}) + m(\bar{x})(x - \bar{x})$ describes how $f(x)$ varies if x varies in an interval around $\bar{x}$. The formula states that $f(x)$ is a straight line with slope $m(\bar{x})$ close to $\bar{x}$.

More generally, if $f(x) = mx + b$ is a linear function, then we can write

$$f(x) = f(\bar{x}) + m\,(x - \bar{x}),$$

since $f(\bar{x}) = b + m\bar{x}$. The coefficient m multiplying the change $x - \bar{x}$ is equal to the derivative of $f(x)$ at $\bar{x}$. In this case, the derivative is constant equal to m for all $\bar{x}$. The change in $f(x)$ is proportional to the change in x with factor of proportionality equal to m:

$$f(x) - f(\bar{x}) = m\,(x - \bar{x}), \tag{23.2}$$

that is if $x \neq \bar{x}$, then the slope m is given by

$$m = \frac{f(x) - f(\bar{x})}{x - \bar{x}} \tag{23.3}$$

We may view the slope m as the change of $f(x)$ per unit change of x, or as the rate of change of $f(x)$ with respect to x.

23.3 Hiking

We now give the above example a different interpretation. Suppose now that x represents time in seconds and $f(x)$ is the distance in meters travelled by a hiker along a hiking path measured from the start at time $x = 0$. According to the above formula, we have $f(x) = 0$ for $x \in [0, 5\,000]$, which means that the trip starts with the hiker at rest at $x = 0$ for $5\,000$ seconds (maybe to fix some malfunctioning equipment). For $x \in [5\,000, 10\,000]$, we have $f(x) = 0.2(x - 5\,000)$ which means that the hiker advances with 0.2 meter per second, that is with the *velocity* 0.2 meters per second. In the time interval $[10\,000, 20\,000]$, we have $f(x) = f(10\,000) + 0.3(x - 10\,000)$, which means that the hiker's velocity is now 0.3 meters per second, and so on.

We note that the slope $m(\bar{x})$ of the straight line $f(x) = f(\bar{x}) + m(\bar{x})(x - \bar{x})$ represents the velocity at $\bar{x}$. We may thus say that the derivative of the distance $f(x)$ with respect to time x, which is the slope $m(\bar{x})$, is equal to the velocity. We will meet the interpretation of the derivative as a velocity again below.

23.4 Definition of the derivative

We shall now seek to define the *derivative* of a given function $f : \mathbb{R} \to \mathbb{R}$ at a given point $\bar{x}$. We shall then follow the idea that if $f(x)$ is particularly well approximated by the linear function $f(\bar{x}) + m\,(x - \bar{x})$ for x close to $\bar{x}$, then the derivative of $f(x)$ at $\bar{x}$ will be equal to m. In other words, the derivative of $f(x)$ at $\bar{x}$ will be equal to the slope m of the approximating linear function $f(\bar{x}) + m\,(x - \bar{x})$. Of course, a key point is to describe how to interpret that the linear function $f(\bar{x}) + m\,(x - \bar{x})$ approximates $f(x)$ "particularly well". We shall see that the natural requirement is to ask that the error is proportional to $|x - \bar{x}|^2$, that is that the error is quadratic in the difference $x - \bar{x}$. Geometrically, this will be the same as asking the straight line $y = f(\bar{x}) + m\,(x - \bar{x})$ to be *tangent* to the graph of $y = f(x)$ at $(\bar{x}, f(\bar{x}))$. We will see that asking the error to be quadratic in $x - \bar{x}$ is just about right. In particular, asking the error to be even smaller, for example proportional to $|x - \bar{x}|^3$, would be to ask for too much.

Before defining the derivative, we back off a little to prepare ourselves and consider different linear approximations $b + m\,(x - \bar{x})$ of the given function $f(x)$ for x close to $\bar{x}$. There are many straight lines that approximate $f(x)$ close to $\bar{x}$. We show some bad approximations and a number of good approximations in Fig. 23.3. On the left, we show some bad linear "approximations" to the function $f(x)$ near $\bar{x}$. On the right, we show some better linear approximations. The question is whether one of the many possible approximate lines is a particularly good choice or not.

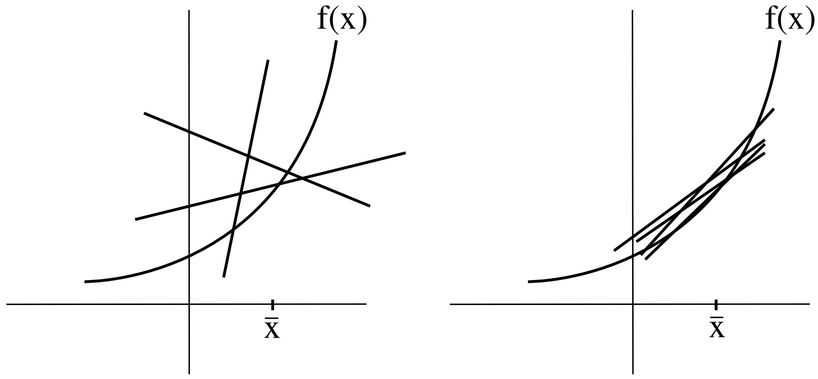


FIGURE 23.3. Linear approximations of $f(x)$ close to $\bar{x}$

We have one piece of information we should use, namely, we know that the value of $f(x)$ at $x = \bar{x}$ is $f(\bar{x})$. So first of all, we only consider lines $b + m\,(x - \bar{x})$ that take on the value $f(\bar{x})$ for $x = \bar{x}$, that is we choose $b = f(\bar{x})$. Such lines are said to *interpolate* $f(x)$ at $\bar{x}$ and thus have an equation of the form

$$y = f(\bar{x}) + m(x - \bar{x}). \tag{23.4}$$

We started this section considering approximations of $f(x)$ of this form. We plot several examples in Fig. 23.4 with different slopes m.

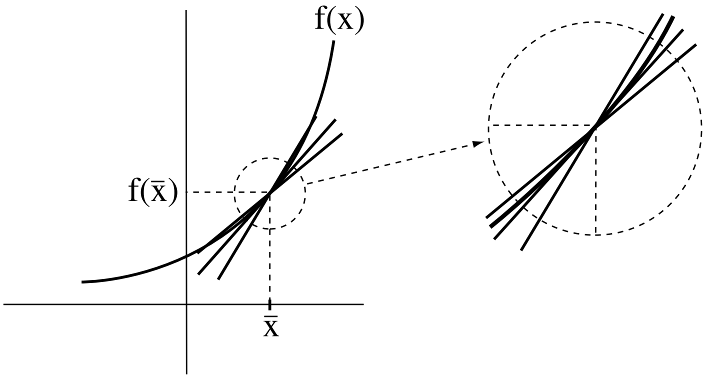


FIGURE 23.4. Linear approximations to a function that pass through the point $(\bar{x}, f(\bar{x}))$. The region near $(\bar{x}, f(\bar{x}))$ has been blown-up on the right.

We now would like to choose the slope m so that $f(x)$ is particularly well approximated by the linear function $f(\bar{x}) + m\,(x - \bar{x})$ for x close to $\bar{x}$. We expect the slope m to depend on $\bar{x}$ and thus we will have $m = m(\bar{x})$.

Out of the three lines plotted in Fig. 23.4 near $(\bar{x}, f(\bar{x}))$, the line in the middle seems to be the best by far. This line is *tangent* to the graph of $f(x)$ at the point $\bar{x}$. The slope of the tangent is characterized by the fact that the error between $f(x)$ and the approximation $f(\bar{x}) + m(\bar{x})(x - \bar{x})$, that is the quantity

$$E_f(x, \bar{x}) = f(x) - \big(f(\bar{x}) + m(\bar{x})(x - \bar{x})\big), \tag{23.5}$$

is particularly small. Since $f(\bar{x}) + m(\bar{x})(x - \bar{x})$ interpolates $f(x)$ at $x = \bar{x}$, we have $E(\bar{x}, \bar{x}) = 0$. Rewriting (23.5) as

$$f(x) = f(\bar{x}) + m(\bar{x})(x - \bar{x}) + E_f(x, \bar{x}),$$

we may view $E_f(x, \bar{x})$ as a correction to the linear approximation $f(\bar{x}) + m(\bar{x})(x - \bar{x})$ of $f(x)$, see Fig. 23.5. It is natural to say that the correction $E_f(x, \bar{x})$ is particularly small if it is much smaller than the term $m(x - \bar{x})$, which represents a linear correction of the constant value $f(\bar{x})$. Thus, $f(\bar{x}) + m(\bar{x})(x - \bar{x})$ is a linear approximation of $f(x)$ close to $\bar{x}$ with zero error for $x = \bar{x}$, and we seek $m(\bar{x})$ so that the correction $E_f(x, \bar{x})$ is small compared to $m(x - \bar{x})$ for x close to $\bar{x}$.

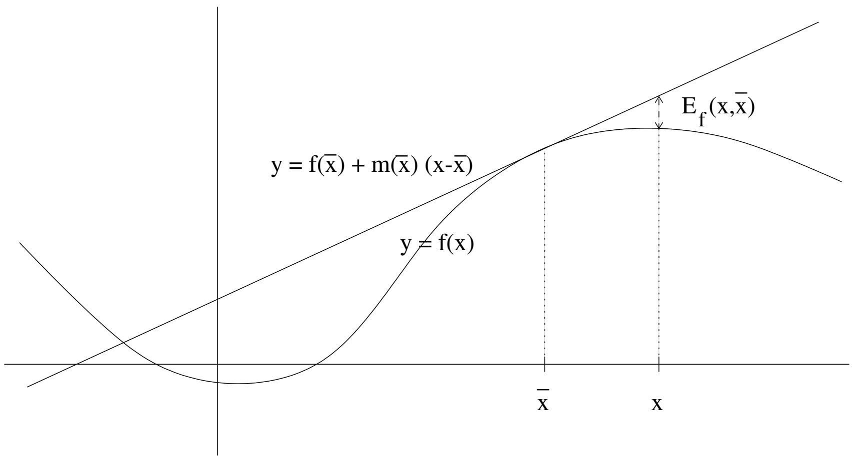


FIGURE 23.5. Graph $y = f(x)$, tangent $y = f(\bar{x}) + m(\bar{x})(x - \bar{x})$ and error $E_f(x, \bar{x})$.

The natural requirement is then to ask that $E_f(x, \bar{x})$ can be bounded by a term which is *quadratic* in $x - \bar{x}$, that is

$$|E_f(x, \bar{x})| \leq K_f(\bar{x})|x - \bar{x}|^2 \quad \text{for } x \text{ close to } \bar{x}, \tag{23.6}$$

where $K_f(\bar{x})$ is a constant. The term $K_f(\bar{x})|x - \bar{x}|^2$ is much smaller than $m(\bar{x})|x - \bar{x}|$, if x is sufficiently close to $\bar{x}$, that is, if the factor $|x - \bar{x}|$ is small enough.

We will say, for short, that an error term $E_f(x,\bar{x})$ is *quadratic* in $x-\bar{x}$ if $E_f(x,\bar{x})$ satisfies the estimate (23.6) for some constant $K_f(\bar{x})$ for x close to $\bar{x}$. We thus seek to choose the slope $m = m(\bar{x})$ so that the error $E_f(x,\bar{x})$ is quadratic in $x-\bar{x}$. The linear function $f(\bar{x}) + m(\bar{x})(x-\bar{x})$ will then be *tangent* to $f(x)$ at $\bar{x}$. We expect the slope of the tangent at $\bar{x}$ to depend on $\bar{x}$, which we indicate by denoting the slope by $m(\bar{x})$.

Now we are in position to define the derivative of $f(x)$ at $\bar{x}$. The function $f(x)$ is said to be *differentiable* at $\bar{x}$ if there are constants $m(\bar{x})$ and $K_f(\bar{x})$ such that for x close to $\bar{x}$,

$$f(x) = f(\bar{x}) + m(\bar{x})(x-\bar{x}) + E_f(x,\bar{x}),$$
$$\text{with } |E_f(x,\bar{x})| \le K_f(\bar{x})|x-\bar{x}|^2. \quad (23.7)$$

We then say that the *derivative* of $f(x)$ at $\bar{x}$ is equal to $m(\bar{x})$, and we denote the derivative by $f'(\bar{x}) = m(\bar{x})$. The derivative $f'(\bar{x})$ of $f(x)$ at $\bar{x}$ is equal to the slope $m(\bar{x})$ of the tangent $y = f(\bar{x}) + m(\bar{x})(x-\bar{x})$ to $f(x)$ at $\bar{x}$. The dependence of $\bar{x}$ is kept in $f'(\bar{x})$.

Recapping our discussion, the equation (23.7) defining the derivative of f at $\bar{x}$ can be thought of as defining a linear approximation

$$f(\bar{x}) + f'(\bar{x})(x-\bar{x}) \approx f(x)$$

for x close to $\bar{x}$ with an error $E_f(x,\bar{x})$ which is quadratic in $x-\bar{x}$. The linear approximation $f(\bar{x}) + f'(\bar{x})(x-\bar{x})$ of $f(x)$ with quadratic error in $x-\bar{x}$, is called the *linearization* of $f(x)$ at $\bar{x}$, and the corresponding $E_f(x,\bar{x})$ is the *linearization error*.

We now compute the derivative of some basic polynomial functions $f(x)$ from the definition of the derivative.

23.5 The derivative of a linear function is constant

If $f(x) = b + mx$ is a linear function with b and m real constants, then

$$f(x) = b + mx = b + m\bar{x} + m(x-\bar{x}) = f(\bar{x}) + m(x-\bar{x}),$$

with the corresponding error function $E_f(x,\bar{x}) = 0$ for all x. We conclude that if $f(x) = b + mx$, then $f'(\bar{x}) = m$. Thus the derivative of a linear function $b + mx$ is constant equal to the slope m. We note that if $m > 0$, then $f(x) = b + mx$ is *increasing* (with increasing x), that is $f(x) > f(\bar{x})$ if $x > \bar{x}$ and $f(x) < f(\bar{x})$ if $x < \bar{x}$. Conversely, if $m < 0$, then $f(x)$ is *decreasing* (with increasing x), that is $f(x) < f(\bar{x})$ if $x > \bar{x}$ and $f(x) > f(\bar{x})$ if $x < \bar{x}$. In particular, for $b = 0$ and $m = 1$ we have

$$\text{if } f(x) = x, \quad \text{then } f'(x) = 1. \quad (23.8)$$

23.6 The derivative of x^2 is $2x$

We now compute the derivative of the quadratic function $f(x) = x^2$ at a point $\bar{x}$. The strategy is to first "extract" the constant value $f(\bar{x})$ from $f(x)$, and a factor $x - \bar{x}$ from the reminder term, to obtain $f(x) = f(\bar{x}) + g(x,\bar{x})(x - \bar{x})$ for some quantity $g(x,\bar{x})$, then to replace $g(x,\bar{x})$ by $g(\bar{x},\bar{x})$ and verify that the resulting error term $E = (g(x,\bar{x}) - g(\bar{x},\bar{x}))(x - \bar{x})$ has the desired property $|E| \leq K|x - \bar{x}|^2$. In the considered case of $f(x) = x^2$ we have

$$x^2 = \bar{x}^2 + (x^2 - \bar{x}^2) = \bar{x}^2 + (x + \bar{x})(x - \bar{x}) = \bar{x}^2 + 2\bar{x}(x - \bar{x}) + (x - \bar{x})^2,$$

that is,

$$f(x) = f(\bar{x}) + 2\bar{x}(x - \bar{x}) + E_f(x,\bar{x}),$$

where $E_f(x,\bar{x}) = (x - \bar{x})^2$, which shows that $f(x) = x^2$ is differentiable for all $\bar{x}$ with $f'(\bar{x}) = 2\bar{x}$, that is, $f'(x) = 2x$ for $x \in \mathbb{R}$. We conclude that, see Fig. 23.6,

$$\text{if } f(x) = x^2, \quad \text{then } f'(x) = 2x. \tag{23.9}$$

An alternative, somewhat shorter route to the linearization formula (23.6) in this case is

$$x^2 = (\bar{x} + (x - \bar{x}))^2 = \bar{x}^2 + 2\bar{x}(x - \bar{x}) + (x - \bar{x})^2,$$

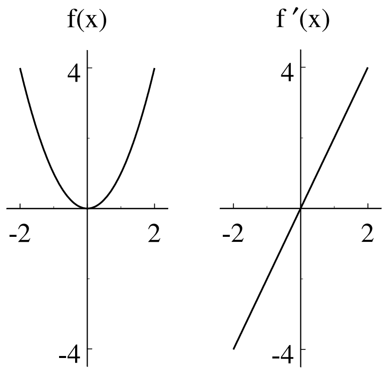


FIGURE 23.6. $f(x) = x^2$ and $f'(x) = 2x$.

We see that x^2 is decreasing for $x < 0$ and increasing for $x > 0$ following the sign of the derivative $f'(x) = 2x$.

Repeating the above calculation with the particular value $\bar{x} = 1$, to get familiar with the argument, we get

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

and thus the derivative of $f(x) = x^2$ at $\bar{x} = 1$ is $f'(1) = 2$. We plot x^2 and $1 + 2(x - 1)$ in Fig. 23.7. We compare some values of the given function x^2

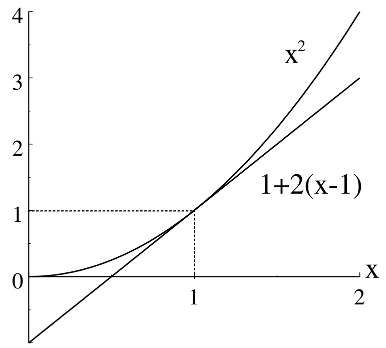


FIGURE 23.7. The linearization $1 + 2(x - 1)$ of x^2 at $\bar{x} = 1$.

to the linear approximation $1 + 2(x - 1)$ along with the error $(x - 1)^2$ in Fig. 23.8

x	$f(x)$	$f(1) + f'(2)(x - 1)$	$E_f(x, 1)$
.7	.49	.4	.09
.8	.64	.6	.04
.9	.81	.8	.01
1.0	1.0	1.0	0.0
1.1	1.21	1.2	.01
1.2	1.44	1.4	.04
1.3	1.69	1.6	.09

FIGURE 23.8. Some values of $f(x) = x^2$, $f(1) + f'(1)(x - 1) = 1 + 2(x - 1)$, and $E_f(x, 1) = (x - 1)^2$.

23.7 The derivative of x^n is nx^{n-1}

We now compute the derivative of the monomial $f(x) = x^n$ at a point $\bar{x}$, where $n \geq 2$ is a natural number. By the Binomial Theorem, generalizing

(23.6), we have

$$x^n = (\bar{x} + x - \bar{x})^n = \bar{x}^n + n\bar{x}^{n-1}(x - \bar{x}) + E_f(x, \bar{x}),$$

where all the terms of the error

$$E_f(x, \bar{x}) = \frac{n(n-1)}{2}\bar{x}^{n-2}(x - \bar{x})^2 + \cdots + (x - \bar{x})^n,$$

contain at least two factors of $(x - \bar{x})$, and thus

$$|E_f(x, \bar{x})| \leq K_f(\bar{x})(x - \bar{x})^2,$$

with $K_f(\bar{x})$ depending on $\bar{x}$, x and n. Clearly, $K_f(\bar{x})$ is bounded by some constant if x and $\bar{x}$ belong to some bounded interval. We conclude that $f'(\bar{x}) = n\bar{x}^{n-1}$ for all $\bar{x}$, that is, $f'(x) = n\bar{x}^{n-1}$ for all x. We summarize:

$$\text{if } f(x) = x^n, \quad \text{then } f'(x) = nx^{n-1}. \tag{23.10}$$

For $n = 2$, we recover the formula $f'(x) = 2x$ if $f(x) = x^2$.

23.8 The derivative of $\frac{1}{x}$ is $-\frac{1}{x^2}$ for $x \neq 0$

We now compute the derivative of the function $f(x) = \frac{1}{x}$ for $x \neq 0$. We have for x close to $\bar{x} \neq 0$,

$$\frac{1}{x} = \frac{1}{\bar{x}} + (\frac{1}{x} - \frac{1}{\bar{x}}) = \frac{1}{\bar{x}} + (-\frac{1}{x\bar{x}})(x - \bar{x}) = \frac{1}{\bar{x}} + (-\frac{1}{\bar{x}^2})(x - \bar{x}) + E$$

where

$$E = (\frac{1}{\bar{x}^2} - \frac{1}{x\bar{x}})(x - \bar{x}) = \frac{1}{x\bar{x}^2}(x - \bar{x})^2,$$

and thus $|E| \leq K|x - \bar{x}|^2$ as desired. We conclude that $f(x) = \frac{1}{x}$ is differentiable at $\bar{x}$ with derivative $f'(\bar{x}) = -\frac{1}{\bar{x}^2}$ for $\bar{x} \neq 0$, that is

$$\text{if } f(x) = \frac{1}{x}, \quad \text{then } f'(\bar{x}) = -\frac{1}{\bar{x}^2} \quad \text{for } \bar{x} \neq 0. \tag{23.11}$$

23.9 The derivative $f'(\bar{x})$ as a function of $\bar{x}$

If a function $f(x)$ is differentiable for all points $\bar{x}$ in an open interval I, then $f(x)$ is said to be *differentiable on* I. The derivative $f'(\bar{x})$ in general varies with $\bar{x}$. We may thus view the derivative $f'(\bar{x})$ of a function $f(x)$, which is differentiable on some interval I, as a function of $\bar{x}$ for $\bar{x} \in I$. We may change the name of the variable $\bar{x}$ and speak about the derivative $f'(x)$ as a function of x. We already took this step above. To a function $f(x)$ that is differentiable on an interval I, we may thus associate the function $f'(x)$ for $x \in I$ that gives the derivative of $f(x)$. We may thus speak of the derivative $f'(x)$ of a differentiable function $f(x)$. For example, the derivative of x^2 is $2x$ and the derivative of x^3 is $3x^2$.

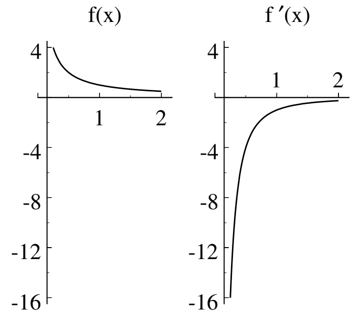


FIGURE 23.9. The function $f(x) = 1/x$ and its derivative $f'(x) = -1/x^2$ for $x > 0$.

23.10 Denoting the derivative of $f(x)$ by $Df(x)$

We also denote the derivative $f'(x)$ of $f(x)$ by $Df(x)$, that is

$$f'(x) = Df(x).$$

Observe that $D(f)$ denotes the domain of f, while $Df(x)$ denotes the derivative of $f(x)$ at x.

We may write the basic formula (23.10) as

$$\text{if } f(x) = x^n, \quad \text{then } f'(x) = Df(x) = nx^{n-1}, \tag{23.12}$$

or

$$Dx^n = nx^{n-1} \quad \text{for } n = 1, 2, ... \tag{23.13}$$

This is one of the most important results of Calculus. We here assume that n is a natural number (including the particular case $n = 0$ if we agree to define $x^0 = 1$ for all x). Below we will extend this formula to n rational (and finally to n real). We recall that we proved above that for $x \neq 0$

$$\text{if } f(x) = \frac{1}{x}, \quad \text{then } f'(x) = Df(x) = -\frac{1}{x^2},$$

corresponding to setting $n = -1$ in (23.12).

EXAMPLE 23.1. Suppose you drive a car along the x-axis and your position at time t measured from the starting point at $t = 0$ is $s(t) = 3\times(2t-t^2)$ miles, where t is measured in hours and the positive direction for s is to the right. Your speed is $s'(t) = 6 - 6t = 6(1 - t)$ miles/hour at time t. Since the derivative is positive for $0 \leq t < 1$, which means that the tangent lines to $s(t)$ have positive slope for $0 \leq t < 1$, the car moves to the right up to $t = 1$. At exactly $t = 1$, you stop the car. If $t > 1$, then the car moves to the left again, because the slopes of the tangents are negative.

23.11 Denoting the derivative of $f(x)$ by $\frac{df}{dx}$

We will also denote the derivative $f'(x)$ of a differentiable function $f(x)$ by

$$\frac{df}{dx} = f'(x) \tag{23.14}$$

We here usually omit the variable x using the notation $\frac{df}{dx}$ and thus write $\frac{df}{dx}$ instead of $\frac{df}{dx}(x)$. Of course the notation $\frac{df}{dx}$ is inspired from (23.22) below, with df corresponding to the f-difference $f(x_i) - f(\bar{x})$ in $f(x)$, and dx corresponding to the x-difference $x_i - \bar{x}$ in x. One may also denote the differentiation operator D in $Df(x)$ alternatively by $\frac{d}{dx}$, and write for example

$$\frac{d}{dx}(x^n) = nx^{n-1} \tag{23.15}$$

We now have three ways of denoting the derivative of a function $f(x)$ with respect to x, namely $f'(x)$, $Df(x)$, and $\frac{df}{dx}$.

Note that using the notation $f'(x)$ and $Df(x)$ for the derivative of a function $f(x)$, it is understood that the derivative is taken with respect to the independent variable x occurring in $f(x)$. This convention is made explicit in the notation $\frac{df}{dx}$. Thus if $f = f(y)$, that is f is a function of the variable y, then $Df = \frac{df}{dy}$, while if $f = f(x)$ then $Df = \frac{df}{dx}$.

23.12 The derivative as a limit of difference quotients

We recall that the function $f(x)$ is differentiable at $\bar{x}$ with derivative $f'(\bar{x})$, if for x in some open interval I containing $\bar{x}$,

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + E_f(x, \bar{x}), \tag{23.16}$$

where

$$|E_f(x, \bar{x})| \leq K_f(\bar{x})|x - \bar{x}|^2, \tag{23.17}$$

and $K_f(\bar{x})$ is a constant. Dividing by $x - \bar{x}$ assuming $x \neq \bar{x}$, we get for $x \in I$,

$$\frac{f(x) - f(\bar{x})}{x - \bar{x}} = f'(\bar{x}) + R_f(x, \bar{x}), \tag{23.18}$$

where

$$R_f(x, \bar{x}) = \frac{E_f(x, \bar{x})}{x - \bar{x}}, \tag{23.19}$$

satisfies

$$|R_f(x, \bar{x})| \leq K_f(\bar{x})|x - \bar{x}| \quad \text{for } x \in I. \tag{23.20}$$

Let now $\{x_i\}_{i=1}^{\infty}$ be a sequence with $\lim_{i\to\infty} x_i = \bar{x}$ with $x_i \in I$ and $x_i \neq \bar{x}$ for all i. There are many such sequences. For example, we may choose $x_i = \bar{x} + i^{-1}$, or $x_i = \bar{x} + 10^{-i}$. From (23.20) it follows that

$$\lim_{i\to\infty} R_f(x_i, \bar{x}) = 0, \tag{23.21}$$

and thus by (23.18) we have

$$f'(\bar{x}) = \lim_{i\to\infty} m_i(\bar{x}), \tag{23.22}$$

where

$$m_i(\bar{x}) = \frac{f(x_i) - f(\bar{x})}{x_i - \bar{x}} \tag{23.23}$$

is the *difference quotient* based on the two distinct points $\bar{x}$ and x_i. The difference quotient $m_i(\bar{x})$ defined by (23.23) is the slope of the *secant line* connecting the points $(\bar{x}, f(\bar{x}))$ and $(x_i, f(x_i))$, see Fig. 23.10, and can be viewed as the *average rate of change* of $f(x)$ between the points $\bar{x}$ and x_i.

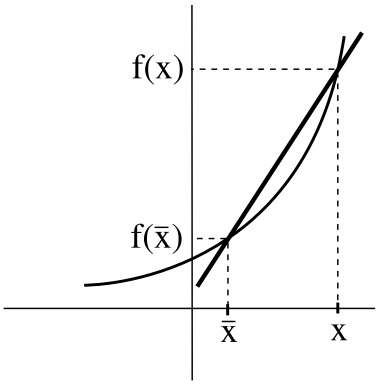


FIGURE 23.10. The secant line joining $(\bar{x}, f(\bar{x}))$ and $(x_i, f(x_i))$.

The formula

$$f'(\bar{x}) = \lim_{i\to\infty} \frac{f(x_i) - f(\bar{x})}{x_i - \bar{x}}, \tag{23.24}$$

expresses the derivative $f'(\bar{x})$ as the limit of the average rate of change of $f(x)$ over intervals $x_i - \bar{x}$, the length of which tend to zero as i tends to infinity. We may thus view $f'(\bar{x})$ as the *local rate of change* of $f(x)$ at $\bar{x}$. If $f(x)$ is tax at income x, then $f'(\bar{x})$ is the marginal tax at $\bar{x}$. If $f(x)$ is a distance and x time, then $f'(\bar{x})$ is the *instantaneous velocity* at time $\bar{x}$.

Alternatively, we may view $f'(\bar{x})$ being the slope of the tangent to $f(x)$ at $x = \bar{x}$ as the limit of the sequence $\{m_i(\bar{x})\}$ of slopes of secants through the points $(\bar{x}, f(\bar{x}))$ and $(x_i, f(x_i))$, where $\{x_i\}_{i=1}^{\infty}$ is a sequence with limit $\bar{x}$. We illustrate in Fig. 23.11.

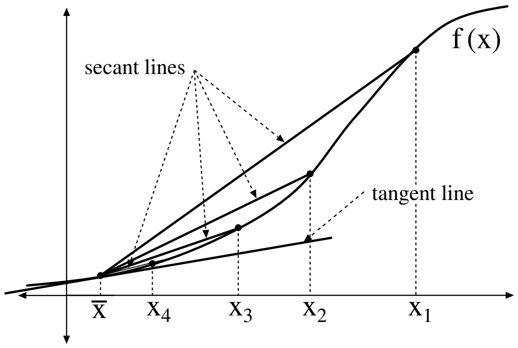


FIGURE 23.11. A sequence of secant lines approaching the tangent line at $\bar{x}$.

EXAMPLE 23.2. Let us now compute the derivative of $f(x) = x^2$ at $\bar{x}$ by using (23.22). Let $x_i = \bar{x} + 1/i$. The slope of the secant line through $(\bar{x}, \bar{x}^2)$ and $(x_i, f(x_i)) = (x_i, x_i^2)$ is

$$m_i(\bar{x}) = \frac{x_i^2 - \bar{x}^2}{x_i - \bar{x}} = \frac{(x_i - \bar{x})(x_i + \bar{x})}{x_i - \bar{x}} = (x_i + \bar{x}).$$

By (23.22), we have

$$f'(\bar{x}) = \lim_{i\to\infty} m_i(\bar{x}) = \lim_{i\to\infty} (2\bar{x} + \frac{1}{i}) = 2\bar{x},$$

and we recover the well known formula $Dx^2 = 2x$.

23.13 How to compute a derivative?

Suppose $f(x)$ is a given function for which we are not able to analytically compute the derivative $f'(\bar{x})$ for a given $\bar{x}$. Note that we were able to carry out the analytical computation above for polynomials, but we gave no strategy to determine the derivative $f'(\bar{x})$ for more general functions $f(x)$. The function $f(x)$ may not be given by any formula at all, and could just be given as a value $f(x)$ for each x determined in some way.

The same problem arises if we want to determine a physical velocity by doing some measurement. For example, if the speed meter of our car is out of function, how can we measure the velocity of the car at some given time $\bar{x}$? Of course the natural thing would be to measure the increment of distance $f(x) - f(\bar{x})$ over some time interval $x - \bar{x}$, where $f(x)$ is the total distance, and then use the quotient $\frac{f(x)-f(\bar{x})}{x-\bar{x}}$, the average velocity over the time interval $(\bar{x}, x)$, as an approximation of the momentary velocity at time $\bar{x}$. But how to choose the length of the time interval $x - \bar{x}$? If we choose $x - \bar{x}$ way too small, then we will not be able to measure any change in position at all, that is we will have $f(x) = f(\bar{x})$, and then conclude zero velocity, while if we take $x - \bar{x}$ too large, the computed average velocity may differ very much from the desired momentary velocity at $\bar{x}$.

We now use analysis to find the right increment $x - \bar{x}$ to use to determine the derivative $f'(\bar{x})$ of a given function $f(x)$ at $\bar{x}$, assuming that the function values $f(x)$ are given with a certain precision. From the definition of $f'(\bar{x})$, we have for x close to $\bar{x}$, $x \neq \bar{x}$,

$$f'(\bar{x}) = \frac{f(x) - f(\bar{x})}{x - \bar{x}} - \frac{E_f(x, \bar{x})}{x - \bar{x}},$$

where

$$|\frac{E_f(x, \bar{x})}{x - \bar{x}}| \leq K_f(\bar{x})|x - \bar{x}|.$$

The difference quotient

$$\frac{f(x) - f(\bar{x})}{x - \bar{x}},$$

may thus be used as an approximation of $f'(\bar{x})$ up to a linearization error of size $K_f(\bar{x})|x - \bar{x}|$.

Suppose now that we know the quantity $f(x) - f(\bar{x})$ up to an error of size δf. We thus assume that we know x and $\bar{x}$ exactly, but there is an error of size δf in the quantity $f(x) - f(\bar{x})$ resulting from errors in the function values $f(x)$ and $f(\bar{x})$ from computation or measurement. We know that frequently the value $f(x)$ for a given x, is known only approximately through computation.

The error δf in $f(x) - f(\bar{x})$ causes an error of size $|\frac{\delta f}{x - \bar{x}}|$ in the difference quotient $\frac{f(x) - f(\bar{x})}{x - \bar{x}}$. We thus have a total error in $f'(\bar{x})$ of size

$$|\frac{\delta f}{x - \bar{x}}| + K_f(\bar{x})|x - \bar{x}|, \tag{23.25}$$

resulting from the error in $f(x) - f(\bar{x})$ and the linearization error. Making the two error contributions equal, which should give the right balance, we get the equation

$$|\frac{\delta f}{x - \bar{x}}| = K_f|x - \bar{x}|,$$

where we write $K_f = K_f(\bar{x})$, from which we compute the "optimal increment"

$$|x - \bar{x}| = \sqrt{\frac{\delta f}{K_f}}. \tag{23.26}$$

If we take $|x - \bar{x}|$ smaller, then the error contribution $|\frac{\delta f}{x - \bar{x}}|$ will dominate and we take $|x - \bar{x}|$ bigger, then the linearization error $K_f(\bar{x})|x - \bar{x}|$ will dominate.

Inserting the optimal increment into (23.25), we get a corresponding "best" error estimate

$$|f'(\bar{x}) - \frac{f(x) - f(\bar{x})}{x - \bar{x}}| \leq 2\sqrt{\delta f}\sqrt{K_f}. \tag{23.27}$$

Contemplating the two resulting formulas (23.26) and (23.27) for the optimal increment and corresponding minimal error in $f'(\bar{x})$, we see that some a priori knowledge of δf and K_f is needed here. If we have no idea of the size of these quantities, we will not know how to choose the increment $x - \bar{x}$ and we will not know anything about the error in the computed derivative. Of course it is in many cases realistic to have an idea of the size of δf, being an error from computation or measurement, but it may be less obvious how to get an idea of the size of K_f. We will return to this question below.

We sum up: Computing an approximation of $f'(\bar{x})$ by using the difference quotient $\frac{f(x)-f(\bar{x})}{x-\bar{x}}$, we should not choose $x - \bar{x}$ too small if there is an error in the quantity $f(x) - f(\bar{x})$. The formula

$$f'(\bar{x}) = \lim_{i\to\infty} \frac{f(x_i) - f(\bar{x})}{x_i - \bar{x}},$$

where $\{x_i\}_{i=1}^{\infty}$ is a sequence with limit $\bar{x}$ and $x_i \neq \bar{x}$, thus must be used with caution. If we examine the cases above where we could compute the derivative analytically, like the case $f(x) = x^2$, we will see that in fact we could divide through by $x_i - \bar{x}$ in the quotient $\frac{f(x_i)-f(\bar{x})}{x_i-\bar{x}}$ and avoid the dangerous appearance of $x_i - \bar{x}$ in the denominator. For example, when computing Dx^2 analytically, we used that

$$\frac{x_i^2 - \bar{x}^2}{x_i - \bar{x}} = \frac{(x_i + \bar{x})(x_i - \bar{x})}{(x_i - \bar{x})} = x_i + \bar{x},$$

from which we could conclude that $Dx^2 = 2x$.

23.14 Uniform differentiability on an interval

We say that the function $f(x)$ is *differentiable* on the interval I if $f(x)$ is differentiable for each $\bar{x} \in I$, that is for $\bar{x} \in I$ there are constants $m(\bar{x})$ and $K_f(\bar{x})$ such that for x close to $\bar{x}$,

$$f(x) = (f(\bar{x}) + m(\bar{x})(x - \bar{x})) + E_f(x, \bar{x})$$
$$|E_f(x, \bar{x})| \leq K_f(\bar{x})|x - \bar{x}|^2.$$

In many cases we can choose one and the same constant $K_f(\bar{x}) = K_f$ for all $\bar{x} \in I$. We may express this by saying the $f(x)$ is uniformly differentiable on I. Allowing also x to vary in I we are led to the following definition, which we will find very useful below: We say that the function $f : I \to \mathbb{R}$ is *uniformly differentiable on the interval* I with derivative $f'(\bar{x})$ at $\bar{x}$, if there is a constant K_f such that for $x, \bar{x} \in I$,

$$f(x) = (f(\bar{x}) + f'(\bar{x})(x - \bar{x})) + E_f(x, \bar{x})$$
$$|E_f(x, \bar{x})| \leq K_f |x - \bar{x}|^2.$$

Observe that the important thing is that K_f here does not depend on $\bar{x}$, but may of course depend on the function f and the interval I.

23.15 Bounded derivative implies Lipschitz continuity

Suppose that $f(x)$ is uniformly differentiable on the interval $I = (a, b)$ and suppose there is a constant L such that for $x \in I$,

$$|f'(x)| \leq L. \tag{23.28}$$

We shall now show that $f(x)$ is Lipschitz continuous on I with Lipschitz constant L, that is we shall show that

$$|f(x) - f(y)| \leq L|x - y| \quad \text{for } x, y \in I. \tag{23.29}$$

This result states something completely obvious: if the absolute value of the maximal rate of change of a function $f(x)$ is bounded by L, then the absolute value of the total change $|f(x) - f(y)|$ is bounded by $L|x - y|$.

If $f(x)$ represents distance, and thus $f'(x)$ velocity, the statement is that if the absolute value of the instantaneous velocity is bounded by L then the absolute value of the change of distance $|f(x) - f(y)|$ is bounded by L times the total time change $|x - y|$. Elementary, my dear Watson!

We shall give a short proof of this result below, when we have some additional machinery available (the Mean Value theorem). We present here a somewhat longer proof.

By assumption we have for $x, y \in I$

$$f(x) = f(y) + f'(y)(x - y) + E_f(x, y),$$

where

$$|E_f(x, y)| \leq K_f |x - y|^2,$$

with K_f a certain constant. We conclude that for $x, y \in I$

$$|f(x) - f(y)| \leq (L + K_f|x - y|)|x - y|,$$

so that for $x, y \in I$,

$$|f(x) - f(y)| \leq \bar{L}|x - y|,$$

where $\bar{L} = L + K(b - a)$. This is almost what we want; the difference is that L is replaced with the somewhat larger Lipschitz constant $\bar{L}$.

If we restrict x and y to a subinterval I_δ of I of length δ, we have

$$|f(x) - f(y)| \leq (L + K\delta)|x - y|$$

By making δ small enough, we can get $L + K\delta$ as close to L as we would like. Let now x and y in I be given and let $x = x_0 < x_1 < \cdots < x_N = y$, where $x_i - x_{i-1} \leq \delta$, see Fig. 23.12

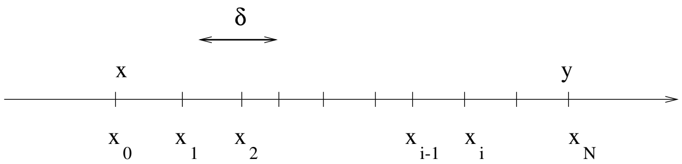


FIGURE 23.12. Subdivision of interval $[x, y]$ into subintervals of length $< \delta$.

We have by the triangle inequality

$$|f(x) - f(y)| = |\sum_{i=1}^{N}(f(x_i) - f(x_{i-1})|$$
$$\leq \sum_{i=1}^{N} |f(x_i) - f(x_{i-1})| \leq (L + K\delta) \sum_{i=1}^{N} |x_i - x_{i-1}|$$
$$= (L + K\delta)|x - y|.$$

Since this inequality holds for any $\delta > 0$, we conclude that indeed

$$|f(x) - f(y)| \leq L|x - y|, \quad \text{for } x, y \in I,$$

which proves the desired result. We summarize in the following theorem which we will use extensively below:

Theorem 23.1 *Suppose that $f(x)$ is uniformly differentiable on the interval $I = (a, b)$ and suppose there is a constant L such that*

$$|f'(x)| \leq L, \quad \textit{for } x \in I.$$

Then $f(x)$ is Lipschitz continuous on I with Lipschitz constant L.

23.16 Connecting to a slightly different definition of the derivative

In many Calculus books the derivative of a function $f : \mathbb{R} \rightarrow \mathbb{R}$ at a point $\bar{x}$ is defined as follows. If the limit

$$\lim_{i\rightarrow\infty} \frac{f(x_i) - f(\bar{x})}{x_i - \bar{x}}, \tag{23.30}$$

does exist for any sequence $\{x_i\}$ with $\lim_{i\rightarrow\infty} x_i = \bar{x}$ (assuming $x_i \neq x$), then we call the (unique) limit the derivative of $f(x)$ at $x = \bar{x}$ and we denote it by $f'(\bar{x})$. We proved in (23.22) that if $f(x)$ is differentiable according to our definition with derivative $f'(\bar{x})$, then

$$f'(\bar{x}) = \lim_{i\rightarrow\infty} \frac{f(x_i) - f(\bar{x})}{x_i - \bar{x}},$$

because we assume that

$$|f'(\bar{x}) - \frac{f(x_i) - f(\bar{x})}{x_i - \bar{x}}| \leq K_f(\bar{x})|x_i - \bar{x}|. \tag{23.31}$$

This means that our definition of derivative is somewhat more demanding than that used in many Calculus books. We assume that the limiting process occurs at a linear rate expressed by (23.31), whereas the definition (23.30) just asks the limit to exist with no rate required (which pleases many mathematicians because of its maximal generality). In most cases, the two concepts agree, but in some very special cases the derivative would exist according to the standard Calculus book definition, but not according to the definition we use. We could naturally relax our definition by relaxing the right hand side bound in (23.31) to $K_f(\bar{x})|x_i - \bar{x}|^\theta$, with some positive constant $\theta < 1$, but the corresponding definition would still be a little stronger than just asking the limit to exist. Using a more demanding definition we focus on normality rather than the extreme or degenerate, which we believe will help the student to approach the new topic. Once the normal situation is understood it may be easier to come to grips with extreme cases.

23.17 Swedenborg

A Swedish counterpart of the Universal Genius Leibniz, together with Newton the Inventor of Calculus, was Emanuel Swedenborg (1688-1772). Swedenborg introduced Calculus to Sweden with independent contributions. Swedenborg produced 150 works on seventeen sciences, was a musician, mining engineer, member of the Swedish parliament, invented a glider, an undersea boat, an ear trumpet for the deaf, a mathematician who wrote the first books in Swedish on algebra and calculus, a physiologist who discovered the function of several areas of the brain and ductless glands, creator of the (at the time) world's largest dry-dock, and suggested the nebula theory of the formation of the planets.

FIGURE 23.13. Emanuel Swedenborg, Swedish Universal Genius, as a young man

Chapter 23 Problems

23.1. Prove directly from the definition that the derivative of x^3 is $3x^2$, and that the derivative of x^4 is $4x^3$.

23.2. Prove directly from the definition that the derivative of the function $f(x) = \sqrt{x} = x^{\frac{1}{2}}$ is equal to $f'(x) = \frac{1}{2}x^{-\frac{1}{2}}$ for $x > 0$. Hint: use that $(\sqrt{x} - \sqrt{\bar{x}})(\sqrt{x} + \sqrt{\bar{x}}) = x - \bar{x}$.

23.3. Compute the derivative of $\sqrt{x}$ numerically for different values of x and study how the error depends on the increment used, and the precision of the computation of $\sqrt{x}$.

23.4. Study the symmetric difference quotient approximation

$$f'(\bar{x}) \approx \frac{f(\bar{x}+h) - f(\bar{x}-h)}{h} \quad h > 0.$$

What is an optimal choice of the increment h, assuming $f(\bar{x} \pm h)$ is not known exactly. Hint: You may find it useful to look ahead into the next chapter (Taylor's formula of order 2).

23.5. Compute the derivative of x^n numerically for different values of x and n and study how the error depends on the increment used.

23.6. Can you compute the derivative of $\sin(x)$ and $\cos(x)$ from the definition?

23.7. Determine the smallest possible Lipschitz constant for the function $f(x) = x^3$ with $D(f) = [1, 4]$.

23.8. (*l'Hopitals rule*). Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be differentiable on an open interval I containing 0, and suppose $f(0) = g(0) = 0$. Prove that

$$\lim_{i \to \infty} \frac{f(x_i)}{g(x_i)} = \frac{f'(0)}{g'(0)}$$

if $g'(0) \neq 0$, where $\{x_i\}_{i=1}^{\infty}$ is a sequence with $\lim_{i \to \infty} x_i = 0$ and $x_i \neq 0$ for all i. This is the famous *l'Hopitals rule*, presented in l'Hopitals book *Analyse de infiment petit* (1713), the first Calculus book! Note that $\frac{f(0)}{g(0)} = \frac{0}{0}$ is not well defined. Hint: Write $f(x_i) = f(0) + f'(x_i)x_i + E_f(x_i, 0)$ et cet.

23.9. Determine $\lim_{i \to \infty} \frac{f(x_i)}{g(x_i)}$, where $f(x) = \sqrt{x} - 1$ and $g(x) = x - 1$, and $\{x_i\}_{i=1}^{\infty}$ is a sequence with $\lim_{i \to \infty} x_i = 1$ and $x_i \neq 1$ for all i. Extend to the case $f(x) = x^r - 1$ with r rational.

24
Differentiation rules

Calculemus. (Leibniz)

When I have followed a line of thought to the end, it often seems so simple that I start to wonder if I have stolen it from someone. (Horace Engdahl)

24.1 Introduction

We now state and prove some rules for computing derivatives of combinations of functions in terms of the derivatives of the functions in the combination. These rules of differentiation form a part of Calculus that can be automated in terms of symbolic manipulation software. In contrast, we will see below that integration, the other basic operation of Calculus, is not open to automatic symbolic manipulation to the same extent. It makes sense that a popular software for symbolic manipulation in Calculus is called *Derive* and not *Integrate*.

The following rules of differentiation are of basic importance and will be used frequently below. They form the very back-bone of symbolic Calculus. Plunging into the proofs we get familiar with different basic aspects of the concept of derivative, and prepare ourselves to write our own version of *Derive*.

24.2 The Linear Combination rule

Suppose that $f(x)$ and $g(x)$ are two functions that are differentiable on an open interval I and let $\bar{x} \in I$. By definition, there are error functions $E_f(x,\bar{x})$ and $E_g(x,\bar{x})$ satisfying for x close to $\bar{x}$,

$$\begin{aligned} f(x) &= f(\bar{x}) + f'(\bar{x})(x-\bar{x}) + E_f(x,\bar{x}), \\ g(x) &= g(\bar{x}) + g'(\bar{x})(x-\bar{x}) + E_g(x,\bar{x}), \end{aligned} \tag{24.1}$$

and

$$|E_f(x,\bar{x})| \leq K_f|x-\bar{x}|^2, \quad |E_g(x,\bar{x})| \leq K_g|x-\bar{x}|^2, \tag{24.2}$$

where K_f and K_g are constants. Addition gives

$$\begin{aligned} f(x)+g(x) = f(\bar{x}) + g(\bar{x}) &+ (f'(\bar{x})+g'(\bar{x}))(x-\bar{x}) \\ &+ E_f(x,\bar{x}) + E_g(x,\bar{x}), \end{aligned}$$

which can be written

$$(f+g)(x) = (f+g)(\bar{x}) + (f'(\bar{x})+g'(\bar{x}))(x-\bar{x}) + E_{f+g}(x,\bar{x}) \tag{24.3}$$

where

$$E_{f+g}(x,\bar{x}) = E_f(x,\bar{x}) + E_g(x,\bar{x}).$$

By (24.2), we have

$$|E_{f+g}(x,\bar{x})| \leq (K_f+K_g)|x-\bar{x}|^2.$$

The formula (24.3) shows that $(f+g)(x)$ is differentiable at $\bar{x}$ and

$$(f+g)'(\bar{x}) = f'(\bar{x}) + g'(\bar{x}). \tag{24.4}$$

Next, multiplying the first line in (24.1) by a constant c, we get

$$(cf)(x) = (cf)(\bar{x}) + cf'(\bar{x})(x-\bar{x}) + cE_f(x,\bar{x}) \tag{24.5}$$

This proves that if $f(x)$ is differentiable at $\bar{x}$, then $(cf)(x)$ is differentiable at $\bar{x}$ and

$$(cf)'(\bar{x}) = cf'(\bar{x}). \tag{24.6}$$

We summarize in

Theorem 24.1 (The Linear Combination rule) *If $f(x)$ and $g(x)$ are differentiable functions on an open interval I and c is a constant, then $(f+g)(x)$ and $(cf)(x)$ are differentiable on I, and for $x \in I$,*

$$(f+g)'(x) = f'(x) + g'(x), \quad \textit{or} \quad D(f+g)(x) = Df(x) + Dg(x), \tag{24.7}$$

and

$$(cf)'(x) = cf'(x), \quad \textit{or} \quad D(cf)(x) = cDf(x). \tag{24.8}$$

Example 24.1.

$$D\big(2x^3 + 4x^5 + \frac{7}{x}\big) = 6x^2 + 20x^4 - \frac{7}{x^2}.$$

Example 24.2. Using the above theorem and the fact that $Dx^i = ix^{i-1}$, we find that the derivative of

$$f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n = \sum_{i=0}^{n} a_ix^i$$

is

$$f'(x) = a_1 + 2a_2x^2 + \cdots + na_nx^{n-1} = \sum_{i=1}^{n} ia_ix^{i-1}.$$

24.3 The Product rule

Multiplying the left and right-hand sides, respectively, of the two equations in (24.1), we obtain

$$\begin{aligned}(fg)(x) = f(x)g(x) = f(\bar{x})g(\bar{x})\\ + f'(\bar{x})g(\bar{x})(x-\bar{x}) + f(\bar{x})g'(\bar{x})(x-\bar{x}) + f'(\bar{x})g'(\bar{x})(x-\bar{x})^2\\ + (g(\bar{x}) + g'(\bar{x})(x-\bar{x}))E_f(x,\bar{x}) + (f(\bar{x})\\ + f'(\bar{x})(x-\bar{x}))E_g(x,\bar{x}) + E_f(x,\bar{x})E_g(x,\bar{x}).\end{aligned}$$

We conclude that

$$(fg)(x) = (fg)(\bar{x}) + \big(f'(\bar{x})g(\bar{x}) + f(\bar{x})g'(\bar{x})\big)(x-\bar{x}) + E_{fg}(x,\bar{x}),$$

where $E_{fg}(x,\bar{x})$ is quadratic in $x-\bar{x}$. We have now proved:

Theorem 24.2 (The Product rule) *If $f(x)$ and $g(x)$ are differentiable on I, then $(fg)(x)$ is differentiable on I and*

$$(fg)'(x) = f(x)g'(x) + f'(x)g(x), \tag{24.9}$$

that is,

$$D(fg)(x) = Df(x)g(x) + f(x)Dg(x), \tag{24.10}$$

Example 24.3.

$$\begin{aligned}&D\left((10+3x^2-x^6)(x-7x^4)\right)\\ &\qquad = (6x-6x^5)(x-7x^4) + (10+3x^2-x^6)(1-28x^3).\end{aligned}$$

24.4 The Chain rule

We shall now compute the derivative of the composite function $(f \circ g)(x) = f(g(x))$ in terms of the derivatives $f'(y) = \frac{df}{dy}$ and $g'(x) = \frac{dg}{dx}$. Suppose then that $g(x)$ is uniformly differentiable on an open interval I, and suppose further that $g(x)$ is Lipschitz continuous on I with Lipschitz constant L_g. Let $\bar{x} \in I$. Suppose next that $f(y)$ is uniformly differentiable on an open interval J containing $\bar{y} = g(\bar{x})$. By definition, there are error functions $E_f(y,\bar{y})$ and $E_g(x,\bar{x})$ satisfying for y close to $\bar{y}$ and x close to $\bar{x}$,

$$\begin{aligned} f(y) &= f(\bar{y}) + f'(\bar{y})(y - \bar{y}) + E_f(y,\bar{y}), \\ g(x) &= g(\bar{x}) + g'(\bar{x})(x - \bar{x}) + E_g(x,\bar{x}), \end{aligned} \tag{24.11}$$

and

$$|E_f(y,\bar{y})| \leq K_f|y - \bar{y}|^2, \quad |E_g(x,\bar{x})| \leq K_g|x - \bar{x}|^2, \tag{24.12}$$

where K_f and K_g are certain constants, independent of y and x, respectively. Further, by assumption

$$|g(x) - g(\bar{x})| \leq L_g|x - \bar{x}|. \tag{24.13}$$

Setting $y = g(x)$ and recalling that $\bar{y} = g(\bar{x})$, we have

$$\begin{aligned} f(g(x)) = f(y) &= f(\bar{y}) + f'(\bar{y})(y - \bar{y}) + E_f(y,\bar{y}) \\ &= f(g(\bar{x})) + f'(g(\bar{x}))(g(x) - g(\bar{x})) + E_f(g(x), g(\bar{x})). \end{aligned}$$

Substituting $g(x) - g(\bar{x}) = g'(\bar{x})(x - \bar{x}) + E_g(x,\bar{x})$, we thus have

$$\begin{aligned} f(g(x)) = f(g(\bar{x})) &+ f'(g(\bar{x}))\, g'(\bar{x})(x - \bar{x}) \\ &+ f'(g(\bar{x}))E_g(x,\bar{x}) + E_f(g(x), g(\bar{x})). \end{aligned}$$

Since (24.12) and (24.13) imply

$$\begin{aligned} |E_f(g(x), g(\bar{x}))| &\leq K_f|g(x) - g(\bar{x})|^2 \leq K_f L_g^2|x - \bar{x}|^2, \\ |f'(g(\bar{x}))\, E_g(x,\bar{x})| &\leq |f'(g(\bar{x}))|\, K_g|x - \bar{x}|^2, \end{aligned}$$

we see that

$$(f \circ g)(x) = (f \circ g)(\bar{x}) + f'(g(\bar{x}))g'(\bar{x})(x - \bar{x}) + E_{f \circ g}(x,\bar{x}),$$

where $E_{f \circ g}(x,\bar{x})$ is quadratic in $x - \bar{x}$. We have now proved:

Theorem 24.3 (The Chain rule) *Assume that $g(x)$ is uniformly differentiable in an open interval I and $g(x)$ is Lipschitz continuous on I. Suppose further that f is uniformly differentiable in an open interval J containing $g(x)$ for x in I. Then the composite function $f(g(x))$ is differentiable on I, and*

$$(f \circ g)'(x) = f'(g(x))g'(x), \quad \textit{for } x \in I, \tag{24.14}$$

or

$$\frac{dh}{dx} = \frac{df}{dy}\frac{dy}{dx}, \tag{24.15}$$

where $h(x) = f(y)$ *and* $y = g(x)$, *that is* $h(x) = f(g(x)) = (f \circ g)(x)$. *An alternative formulation si*

$$D(f(g(x)) = Df(g(x))Dg(x), \tag{24.16}$$

where $Df = \frac{df}{dy}$.

EXAMPLE 24.4. Let $f(y) = y^5$ and $y = g(x) = 9-8x$, so that $f(g(x)) = (f \circ g)(x) = (9-8x)^5$. We have $f'(y) = 5y^4$ and $g'(x) = -8$, and thus

$$D((9-8x)^5) = 5y^4\, g'(x) = 5(9-8x)^4\,(-8) = -40(9-8x)^4.$$

EXAMPLE 24.5.

$$\begin{aligned} D\big(7x^3+4x+6\big)^{18} &= 18\big(7x^3+4x+6\big)^{17} D(7x^3+4x+6) \\ &= 18\big(7x^3+4x+6\big)^{17}(21x^2+4). \end{aligned}$$

EXAMPLE 24.6. Consider the composite function $f(g(x))$ with $f(y) = 1/y$, that is the function $h(x) = \frac{1}{g(x)}$, where $g(x)$ is a given function with $g(x) \neq 0$. Since $Df(y) = -\frac{1}{y^2}$ we have using the Chain rule

$$Dh(x) = D\frac{1}{g(x)} = \frac{-1}{(g(x))^2}\, g'(x) = \frac{-g'(x)}{g(x)^2}, \tag{24.17}$$

as long as $g(x)$ is differentiable and $g(x) \neq 0$.

EXAMPLE 24.7. Using Example 24.6 and the Chain Rule, we get for $n \geq 1$

$$\begin{aligned} \frac{d}{dx}x^{-n} = \frac{d}{dx}\left(\frac{1}{x^n}\right) &= \frac{-1}{(x^n)^2}\frac{d}{dx}x^n \\ &= \frac{-1}{x^{2n}} \times nx^{n-1} = -nx^{-n-1}. \end{aligned}$$

This extends the formula $Dx^m = mx^{m-1}$ to negative integers $m = -1, -2, ...$

24.5 The Quotient rule

Let $f(x)$ and $g(x)$ be differentiable on I and consider the problem of computing the deriviative of $(\frac{f}{g})(x) = \frac{f(x)}{g(x)}$ at $\bar{x}$. Applying the Product rule to $f(x)\,\frac{1}{g(x)} = \frac{f(x)}{g(x)}$, and using (24.17), we find that

$$\left(\frac{f}{g}\right)'(\bar{x}) = f'(\bar{x})\frac{1}{g(\bar{x})} + f(\bar{x})\frac{-g'(\bar{x})}{g(\bar{x})^2} = \frac{f'(\bar{x})g(\bar{x}) - f(\bar{x})g'(\bar{x})}{g(\bar{x})^2},$$

if $g(\bar{x}) \neq 0$, and we have thus proved:

Theorem 24.4 (The Quotient rule) *Assume that $f(x)$ and $g(x)$ are differentiable functions on the open interval I. Then for $x \in I$, we have*

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2},$$

provided $g(x) \neq 0$.

EXAMPLE 24.8.

$$D\left(\frac{3x+4}{x^2-1}\right) = \frac{3 \times (x^2-1) - (3x+4) \times 2x}{(x^2-1)^2}.$$

EXAMPLE 24.9.

$$\begin{aligned}
&\frac{d}{dx}\left(\frac{x^3+x}{(8-x)^6}\right)^9 \\
&\quad = 9\left(\frac{x^3+x}{(8-x)^6}\right)^8 \frac{d}{dx}\left(\frac{x^3+x}{(8-x)^6}\right) \\
&\quad = 9\left(\frac{x^3+x}{(8-x)^6}\right)^8 \frac{(8-x)^6 \frac{d}{dx}(x^3+x) - (x^3+x)\frac{d}{dx}(8-x)^6}{\left((8-x)^6\right)^2} \\
&\quad = 9\left(\frac{x^3+x}{(8-x)^6}\right)^8 \frac{(8-x)^6(3x^2+1) - (x^3+x)6(8-x)^5 \times -1}{(8-x)^{12}}.
\end{aligned}$$

EXAMPLE 24.10. The Chain rule can also be used recursively:

$$\begin{aligned}
&\frac{d}{dx}((((1-x)^2+1)^3+2)^4+3)^5 \\
&\quad = 5((((1-x)^2+1)^3+2)^4+3)^4 \times 4(((1-x)^2+1)^3+2)^3 \\
&\qquad\qquad \times 3((1-x)^2+1)^2 \times 2(1-x) \times (-1).
\end{aligned}$$

24.6 Derivatives of derivatives: $f^{(n)} = D^n f = \frac{d^n f}{dx^n}$

Let $f(x)$ be a function with derivative $f'(x)$. Since $f'(x)$ is a function, it may also be differentiable with a derivative which would describe how quickly the rate of change of f is changing at each point x. The derivative of the derivative $f'(x)$ of $f(x)$ is called the *second derivative* of $f(x)$ and is denoted by

$$f''(x) = D^2 f(x) = \frac{d^2 f}{dx^2} = (f')'(x).$$

EXAMPLE 24.11. For $f(x) = x^2$, $f'(x) = 2x$ and $f''(x) = 2$.

EXAMPLE 24.12. For $f(x) = 1/x$, $f'(x) = -1/x^2 = -x^{-2}$ and $f''(x) = -(-2)x^{-3} = 2/x^3$.

We can continue taking the derivative of the second derivative and get a third derivative:

$$f'''(x) = D^3 f(x) = \frac{d^3 f}{dx^3} = (f'')'(x)$$

as long as the functions are differentiable. We can recursively define the derivative $f^{(n)} = D^n f$ of f order n by

$$f^{(n)}(x) = D^n f(x) = \frac{d^n f}{dx^n} = (f^{(n-1)})'(x) = D(D^{n-1} f)(x),$$

where $f'(x) = f^{(1)}(x) = Df(x)$, $f''(x) = f^{(2)}(x) = D^2 f(x)$, and so on.

The derivative of distance with respect to time is velocity. The derivative of velocity with respect to time is called *acceleration.* Velocity indicates how quickly the position of an object is changing with time and acceleration indicates how quickly the object is speeding up or slowing down (changing velocity) with respect to time.

EXAMPLE 24.13. If $f(x) = x^4$, then $Df(x) = 4x^3$, $D^2 f(x) = 12x^2$, $D^3 f(x) = 24x$, $D^4 f(x) = 24$ and $D^5 f(x) \equiv 0$.

EXAMPLE 24.14. The $n + 1$'st derivative of a polynomial of degree n is zero.

EXAMPLE 24.15. If $f(x) = 1/x$, then

$$f(x) = x^{-1},\ Df(x) = -1 \times x^{-2},\ D^2 f(x) = 2 \times x^{-3},\ D^3 f(x) = -6 \times x^{-4}$$

$$\vdots$$

$$D^n f(x) = (-1)^n \times 1 \times 2 \times 3 \times \cdots \times n x^{-n-1} = (-1)^n n! x^{-n-1}.$$

24.7 One-sided derivatives

We can also define *differentiability from the right* at a point $\bar{x}$ of a function $f(x)$. The definition is the same as that used above with the restriction that $x \geq \bar{x}$. More precisely, the function $f : J \to \mathbb{R}$, where $J = [\bar{x}, b)$ and $b > \bar{x}$, is said to be *differentiable from the right* at $\bar{x}$ if there are constants $m(\bar{x})$ and $K_f(\bar{x})$ such that for $x \in [\bar{x}, b)$

$$|f(x) - (f(\bar{x}) + m(\bar{x})(x - \bar{x}))| \leq K_f(\bar{x})|x - \bar{x}|^2. \qquad (24.18)$$

We then say that the *right-hand derivative* of $f(x)$ at $\bar{x}$ is equal to $m(\bar{x})$, and we denote the right-hand derivative by $f'_+(\bar{x}) = m(\bar{x})$.

We define the left-hand derivative $f'_-(\bar{x}) = m(\bar{x})$, analogously restricting $x \leq \bar{x}$. In both cases, we are simply requiring that the linearization estimate holds for x on one side of $\bar{x}$.

EXAMPLE 24.16. The function $f(x) = |x|$ is differentiable for $\bar{x} \neq 0$ with derivative $f'(\bar{x}) = 1$ if $\bar{x} > 0$ and $f'(\bar{x}) = -1$ if $\bar{x} < 0$. The function $f(x) = |x|$ is differentiable from the right at $\bar{x} = 0$ with derivative $f'_+(0) = 1$, and differentiable form the left at $\bar{x} = 0$ with derivative $f'_-(0) = -1$.

We say that $f : [a, b] \to \mathbb{R}$ is *differentiable on the closed interval* $[a, b]$, if $f(x)$ is differentiable on the open interval (a, b), and is differentiable from the right at a, and differentiable from the left at b. The definition extends in the obvious way to half-open/half-closed intervals $(a, b]$ and $[a, b)$. If f is either differentiable or is differentiable from the right and/or the left at every point in an interval, then we say that f is *piecewise differentiable* on the interval.

EXAMPLE 24.17. The function $|x|$ is piecewise differentiable on $\mathbb{R}$. The function $1/x$ is differentiable on $(0, \infty)$ but not differentiable on $[0, \infty)$.

24.8 Quadratic approximation: Taylor's formula of order two

For a differentiable function $f(x)$, we figured out how to compute a best linear approximation for x close to $\bar{x}$, namely

$$f(x) \approx f(\bar{x}) + f'(\bar{x})(x - \bar{x})$$

with an error quadratic in $x - \bar{x}$. In some situations, we might require more accuracy from an approximation than is possible to get using a linear function. The natural generalization is to look for a "best" quadratic approximation of the form

$$f(x) = f(\bar{x}) + m_1(\bar{x})(x - \bar{x}) + m_2(\bar{x})(x - \bar{x})^2 + E_f(x, \bar{x}), \tag{24.19}$$

for x close to $\bar{x}$, where $m_1(\bar{x})$ and $m_2(\bar{x})$ are constants and now the error function $E_f(x, \bar{x})$ is *cubic* in $x - \bar{x}$, that is

$$|E_f(x, \bar{x})| \leq K_f(\bar{x})|x - \bar{x}|^3, \tag{24.20}$$

with $K_f(\bar{x})$ a constant. Of course, for $|x - \bar{x}|$ small, $K_f(\bar{x})|x - \bar{x}|^3$ is much smaller than both $m_1(\bar{x})(x - \bar{x})$ or $m_2(\bar{x})(x - \bar{x})^2$, unless $m_1(\bar{x})$ and $m_2(\bar{x})$ happen to be zero, of course.

Now, if (24.19) holds for x close to $\bar{x}$, then $m_1(\bar{x}) = f'(\bar{x})$, since $m_2(x-\bar{x})^2 + E_f(x,\bar{x})$ is quadratic in $x-\bar{x}$. If (24.19) holds, we thus have

$$f(x) = f(\bar{x}) + f'(\bar{x})(x-\bar{x}) + m_2(\bar{x})(x-\bar{x})^2 + E_f(x,\bar{x}). \tag{24.21}$$

Let us next try to determine the constant $m_2(\bar{x})$. To this end we differentiate the relation (24.19) with respect to x to get

$$f'(x) = f'(\bar{x}) + 2m_2(\bar{x})(x-\bar{x}) + \frac{d}{dx}E_f(x,\bar{x}). \tag{24.22}$$

Let us now assume that for x close to $\bar{x}$

$$|\frac{d}{dx}E_f(x,\bar{x})| \leq M_f(\bar{x})|x-\bar{x}|^2, \tag{24.23}$$

for some constant $M_f(\bar{x})$. The principle is that taking the derivative brings down the power of $|x-\bar{x}|$ one step from 3 to 2. We shall meet this phenomenon many times below. From (24.23) it would then follow by the definition of $f''(\bar{x})$, that $f''(\bar{x}) = (f')'(\bar{x}) = 2m_2(\bar{x})$, that is

$$m_2(\bar{x}) = \frac{1}{2}f''(\bar{x}).$$

We would thus arrive at an approximation formula of the form

$$f(x) = f(\bar{x}) + f'(\bar{x})(x-\bar{x}) + \frac{1}{2}f''(\bar{x})(x-\bar{x})^2 + E_f(x,\bar{x}), \tag{24.24}$$

for x close to $\bar{x}$, where $|E_f(x,\bar{x})| \leq K_f(\bar{x})|x-\bar{x}|^3$ with $K_f(\bar{x})$ a constant.

Example 24.18. Consider the function $f(x) = \frac{1}{x}$ for x close to $\bar{x} = 1$. We shall use the fact that if $y \neq -1$, then

$$\frac{1}{1+y} = 1 - y\frac{1}{1+y}$$

which is readily verified by multiplying by $1+y$, and thus

$$\frac{1}{1+y} = 1 - y\frac{1}{1+y} = 1 - y(1 - y\frac{1}{1+y})$$

$$= 1 - y + y^2\frac{1}{1+y} = 1 - y + y^2 - y^3\frac{1}{1+y}.$$

Choosing $y = x-1$, we get

$$\frac{1}{x} = \frac{1}{1+(x-1)} = 1 - (x-1) + (x-1)^2 - \frac{(x-1)^3}{1+(x-1)}, \tag{24.25}$$

and we see that the quadratic polynomial

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

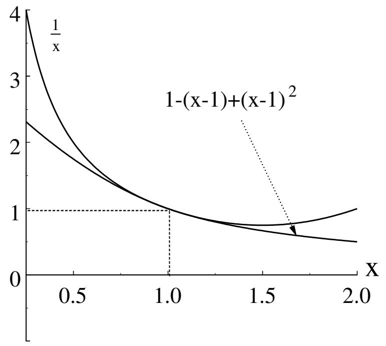


FIGURE 24.1. The quadratic approximation $1-(x-1)+(x-1)^2$ of $1/x$ near $\bar{x}=1$

x	$1/x$	$1-(x-1)+(x-1)^2$	$E_f(x,1)$
.7	1.428571	1.39	.038571
.8	1.25	1.22	.03
.9	1.111111	1.11	.00111
1.0	1.0	1.0	0.0
1.1	.909090	.91	.000909
1.2	.833333	.84	.00666
1.3	.769230	.79	.02077

FIGURE 24.2. Some values of $f(x) = 1/x$, the quadratic approximation $1-(x-1)+(x-1)^2$, and the error $E_f(x,1)$.

approximates $\frac{1}{x}$ for x close to $\bar{x} = 1$ with an error, which is cubic in $x - \bar{x}$. As a consequence of the expansion, we have that $f(1) = 1$, $f'(1) = -1$ and $f''(1) = 2$. We plot the approximation in Fig. 24.1 and list some values of the approximation in Fig. 24.2.

Below we will prove under the name of *Taylor's theorem*, that if the function $f(x)$ is three times differentiable with $|f^{(3)}(x)| \leq 6K_f(\bar{x})$ for x close to $\bar{x}$, where $K_f(\bar{x})$ is a constant, then for x close to $\bar{x}$,

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + \frac{1}{2}f''(\bar{x})(x - \bar{x})^2 + E_f(x, \bar{x}), \qquad (24.26)$$

where the error function $E_f(x, \bar{x})$ is cubic in $x - \bar{x}$, more precisely,

$$|E_f(x, \bar{x})| \leq K_f(\bar{x})|x - \bar{x}|^3, \quad \text{for } x \text{ close to } \bar{x}. \qquad (24.27)$$

Further, $\frac{d}{dx}E_f(x, \bar{x})$ is quadratic in $x - \bar{x}$. Taylor's theorem thus gives an answer to the problem of quadratic approximation formulated in (24.19).

24.9 The derivative of an inverse function

Let $f : (a, b) \to \mathbb{R}$ be differentiable at $\bar{x} \in (a, b)$, so that for x close to $\bar{x}$

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + E_f(x, \bar{x}), \qquad (24.28)$$

where $|E_f(x, \bar{x})| \leq K_f(\bar{x})(x - \bar{x})^2$ with $K_f(\bar{x})$ a constant. Suppose that $f'(\bar{x}) \neq 0$ so that $f(x)$ is strictly increasing or decreasing for x close to $\bar{x}$, and thus the equation $y = f(x)$ has a unique solution x for y close to $\bar{y} = f(\bar{x})$. This defines x as a function of y, and this function is said to be the *inverse* of the function $y = f(x)$ and is denoted by $x = f^{-1}(y)$, see (24.3).

Can we compute the derivative of the function $x = f^{-1}(y)$ with respect to y close to $\bar{y} = f(\bar{x})$? Rewriting (24.28), we have

$$y = \bar{y} + f'(\bar{x})(f^{-1}(y) - f^{-1}(\bar{y})) + E_f(f^{-1}(y), f^{-1}(\bar{y})),$$

that is

$$f^{-1}(y) = f^{-1}(\bar{y}) + \frac{1}{f'(\bar{x})}(y - \bar{y}) - \frac{1}{f'(\bar{x})}E_f(f^{-1}(y), f^{-1}(\bar{y})), \qquad (24.29)$$

Suppose now that f^{-1} is Lipschitz continuous in an open interval J around $\bar{y}$, so that

$$|f^{-1}(y) - f^{-1}(\bar{y})| \leq L_{f^{-1}}|y - \bar{y}| \quad \text{for } y \in J.$$

Then for y close to $\bar{y}$,

$$|\frac{1}{f'(\bar{x})}E_f(f^{-1}(y), f^{-1}(\bar{y}))| \leq \frac{1}{|f'(\bar{x})|}K_f(\bar{x})(L_{f^{-1}})^2|y - \bar{y}|^2,$$

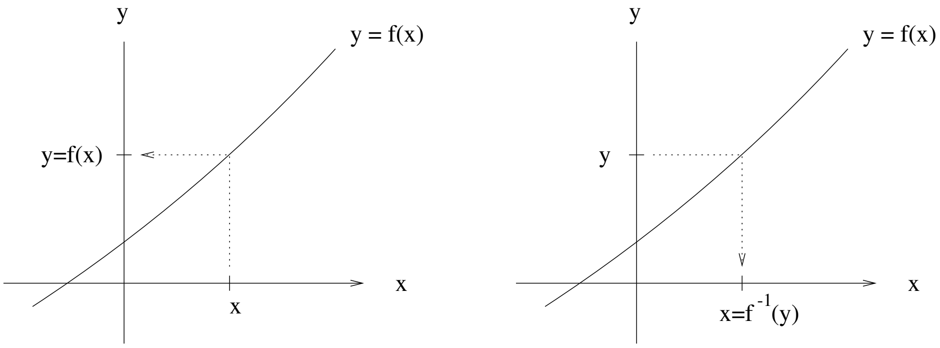


FIGURE 24.3. The function $y = f(x)$ and its inverse $x = f^{-1}(y)$

which proves by (24.29) that the derivative $Df^{-1}(\bar{y})$ of $f^{-1}(y)$ with respect to y at $\bar{y}$ is equal to $\frac{1}{f'(\bar{x})}$, that is

$$Df^{-1}(\bar{y}) = \frac{1}{f'(\bar{x})}, \tag{24.30}$$

where $\bar{y} = f(\bar{x})$. We summarize:

Theorem 24.5 *If $y = f(x)$ is differentiable at $\bar{x}$ with respect to x with $f'(\bar{x}) \neq 0$, then the inverse function $x = f^{-1}(y)$ is differentiable with respect to y at $\bar{y} = f(\bar{x})$ with derivative $Df^{-1}(\bar{y}) = \frac{1}{f'(\bar{x})}$.*

EXAMPLE 24.19. The inverse of the function $y = f(x) = x^2$ for $x > 0$ is the function $x = f^{-1}(y) = \sqrt{y}$ defined for $y > 0$. It follows that $D\sqrt{y} = \frac{1}{f'(x)} = \frac{1}{2x} = \frac{1}{2\sqrt{y}}$. Changing notation from y to x, we thus have for $x > 0$,

$$\frac{d}{dx}\sqrt{x} = D\sqrt{x} = \frac{1}{2\sqrt{x}}, \quad \text{or} \quad Dx^{\frac{1}{2}} = \frac{1}{2}x^{-\frac{1}{2}}. \tag{24.31}$$

24.10 Implicit differentiation

We give an example of a technique called *implicit differentiation* to compute the derivative of the function $x^{\frac{p}{q}}$, where p and q are integers with $q \neq 0$, and $x > 0$. We know that the function $y = x^{\frac{p}{q}}$ is the unique solution of the equation $y^q = x^p$ in y for a given $x > 0$. We can thus view y as a function of x and write $y(x) = x^{\frac{p}{q}}$, and we have

$$(y(x))^q = x^p \quad \text{for } x > 0. \tag{24.32}$$

Assuming $y(x)$ to be differentiable with respect to x with derivative $y'(x)$, we would get differentiating both sides of (24.32) with respect to x, and using the Chain Rule on the left hand side:

$$q(y(x))^{q-1}y'(x) = px^{p-1}$$

from which we deduce inserting that $y(x) = x^{\frac{p}{q}}$,

$$y'(x) = \frac{p}{q}x^{-\frac{p}{q}(q-1)}x^{p-1} = \frac{p}{q}x^{\frac{p}{q}-1}.$$

We conclude that

$$Dx^r = rx^{r-1} \quad \text{for } r \text{ rational, and } x > 0, \tag{24.33}$$

using the computation as an indication that the derivative indeed exists.

To connect with the previous section, note that if $y = f(x)$ has an inverse function $x = f^{-1}(y)$, then differentiating both sides of $x = f^{-1}(y)$ with respect to x, considering $y = y(x) = f(x)$ as a function of x, we get with $D = \frac{d}{dy}$

$$1 = Df^{-1}(y)f'(x)$$

which gives the formula (24.30).

24.11 Partial derivatives

We now have gained some experience of the concept of derivative of a real-valued function $f : \mathbb{R} \to \mathbb{R}$ of one real variable x. Below we shall consider real-valued functions *several real variables*, and we are then led to the concept of *partial derivative*. We give here a first glimpse, and consider a real-valued function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ of two real variables, that is for each $x_1 \in \mathbb{R}$ and $x_2 \in \mathbb{R}$, we are given a real number $f(x_1, x_2)$. For example,

$$f(x_1, x_2) = 15x_1 + 3x_2, \tag{24.34}$$

represents the total cost in the Dinner Soup/Ice Cream model, with x_1 representing the amount of meat and x_2 the amount of ice-cream. To compute the *partial derivative* of the function $f(x_1, x_2) = 15x_1 + 3x_2$ with respect to x_1, we keep the variable x_2 constant and compute the derivative of the function $f_1(x_1) = f(x_1, x_2)$ as a function of x_1, and obtain $\frac{df_1}{dx_1} = 15$, and we write

$$\frac{\partial f}{\partial x_1} = 15$$

which is the *partial derivative of* $f(x_1, x_2)$ *with respect to* x_1. Similarly, to compute the *partial derivative* of the function $f(x_1, x_2) = 15x_1 + 3x_2$ with respect to x_2, we keep the variable x_1 constant and compute the derivative of the function $f_2(x_2) = f(x_1, x_2)$ as a function of x_2, and obtain $\frac{df_2}{dx_2} = 3$, and we write

$$\frac{\partial f}{\partial x_2} = 3.$$

Obviously, $\frac{\partial f}{\partial x_1}$ represents the cost of increasing the amount of meat one unit, and $\frac{\partial f}{\partial x_2} = 3$ represents the cost of increasing the amount of ice cream one unit. The *marginal cost* of meat is thus $\frac{\partial f}{\partial x_1} = 15$ and that of ice cream $\frac{\partial f}{\partial x_2} = 3$.

EXAMPLE 24.20. Suppose $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is given by $f(x_1, x_2) = x_1^2 + x_2^3 + x_1 x_2$. We compute

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = 2x_1 + x_2, \quad \frac{\partial f}{\partial x_2}(x_1, x_2) = 3x_2^2 + x_1,$$

where we follow the principle just explained: to compute $\frac{\partial f}{\partial x_1}$, keep x_2 constant and differentiate with respect to x_1, and to compute $\frac{\partial f}{\partial x_2}$, keep x_1 constant and differentiate with respect to x_2.

More generally, we may in a natural way extend the concept of differentiability of a real-valued function $f(x)$ of one real variable x to differentiability of a real valued function $f(x_1, x_2)$ of two real variables x_1 and x_2 as follows: We say that function $f(x_1, x_2)$ is *differentiable* at $\bar{x} = (\bar{x}_1, \bar{x}_2)$ if there are constants $m_1(\bar{x}_1, \bar{x}_2)$, $m_2(\bar{x}_1, \bar{x}_2)$ and $K_f(\bar{x}_1, \bar{x}_2)$, such that for (x_1, x_2) close to $(\bar{x}_1, \bar{x}_2)$,

$$f(x_1, x_2) = f(\bar{x}_1, \bar{x}_2) + m_1(\bar{x}_1, \bar{x}_2)(x_1 - \bar{x}_1) + m_2(\bar{x}_1, \bar{x}_2)(x_2 - \bar{x}_2) + E_f(x, \bar{x}),$$

where

$$|E_f(x, \bar{x})| \le K_f(\bar{x}_1, \bar{x}_2)((x_1 - \bar{x}_1)^2 + (x_2 - \bar{x}_2)^2).$$

Note that

$$f(\bar{x}_1, \bar{x}_2) + m_1(\bar{x}_1, \bar{x}_2)(x_1 - \bar{x}_1) + m_2(\bar{x}_1, \bar{x}_2)(x_2 - \bar{x}_2)$$

is a linear approximation to $f(x)$ with quadratic error, the graph of which represents the *tangent plane* to $f(x)$ at $\bar{x}$.

Letting x_2 be constant equal to $\bar{x}_2$, we see that the *partial derivative* of $f(x_1, x_2)$ at $(\bar{x}_1, \bar{x}_2)$ with respect to x_1 is equal to $m_1(\bar{x}_1, \bar{x}_2)$, and we denote this derivative by

$$\frac{\partial f}{\partial x_1}(\bar{x}_1, \bar{x}_2) = m_1(\bar{x}_1, \bar{x}_2).$$

Similarly, we say that the *partial derivative* of $f(x)$ at $\bar{x}$ with respect to x_2 is equal to $m_2(\bar{x}_1, \bar{x}_2)$ and denote this derivative by $\frac{\partial f}{\partial x_2}(\bar{x}_1, \bar{x}_2) = m_2(\bar{x}_1, \bar{x}_2)$.

These ideas extend in a natural way to real-valued functions $f(x_1, ...x_d)$ of d real variables x_1,...,x_d, and we can speak about (and compute) partial derivatives of $f(x_1, ...x_d)$ with respect to x_1,...,x_d following the same basic idea. To compute the partial derivative $\frac{\partial f}{\partial x_j}$ with respect to x_j for some $j = 1, ..., d$, we keep all variables but x_j constant and compute the usual derivative with respect to x_j. We shall return below to the concept of partial derivative below, and through massive experience learn that it plays a basic role in mathematical modeling.

24.12 A sum up so far

We have proved above that

$$Dx^n = \frac{d}{dx}x^n = nx^{n-1} \quad \text{for } n \text{ integer and } x \neq 0,$$
$$Dx^r = \frac{d}{dx}x^r = rx^{r-1} \quad \text{for } r \text{ rational and } x > 0.$$

We have also proved rules for how to differentiate linear combinations, products, quotients, compositions, and inverses of differentiable functions. This is just about all so far. We lack in particular answers to the following questions:

- What function $u(x)$ satisfies $u'(x) = \frac{1}{x}$?
- What is the derivative of the function a^x, where $a > 0$ is a constant?

Chapter 24 Problems

24.1. Construct and differentiate functions obtained by combining functions of the form x^r using linear combinations, products, quotients, compositions, and taking inverses. For example, functions like

$$\sqrt{x^{11} + \sqrt{\frac{x^{111}}{x^{-1.1} + x^{1.1}}}}.$$

24.2. Compute the partial derivatives of the function $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ defined by $f(x_1, x_2) = x_1^2 + x_2^4$.

24.3. We have defined 2^x for x rational. Let us try to compute the derivative $D2^x = \frac{d}{dx}2^x$ with respect to x at $x = 0$. We are then led to study the quotient

$$q_n = \frac{2^{\frac{1}{n}} - 1}{\frac{1}{n}}$$

as n tends to infinity. (a) Do this experimentally using the computer. Note that $2^{\frac{1}{n}} = 1 + \frac{q_n}{n}$, and thus we seek q_n so that $(1 + \frac{q_n}{n})^n = 2$. Compare with the experience concerning $(1 + \frac{1}{n})^n$ in Chapter A Very Short Course in Calculus.

24.4. Suppose you know how to compute the derivative of 2^x at $x = 0$. What is the derivative then at $x \neq 0$? Hint: $2^{x+\frac{1}{n}} = 2^x 2^{\frac{1}{n}}$.

24.5. Consider the function $f : (0, 2) \to \mathbb{R}$ defined by $f(x) = (1 + x^4)^{-1}$ for $0 < x < 1$, $f(x) = ax + b$ for $1 \leq x < 2$, where $a, b \in \mathbb{R}$ are constants. For what values of a and b is this function (i) Lipschitz continuous on $(0, 2)$, (ii) differentiable on $(0, 2)$?

24.6. Compute the partial derivatives of the function $f : \mathbb{R}^3 \to \mathbb{R}$ given by $f(x_1, x_2, x_3) = 2x_1^2 x_3 + 5x_2^3 x_3^4$.

25
Newton's method

Brains first and then Hard Work. (The House at Pooh Corner, Milne)

25.1 Introduction

As a basic application of the derivative, we study *Newton's method* for computing roots of an equation $f(x) = 0$. Newton's method is one of the corner-stones of constructive mathematics. As a preparation we start out using the concept of derivative to analyze the convergence of Fixed Point Iteration.

25.2 Convergence of fixed point iteration

Let $g : I \to I$ be uniformly differentiable on an interval $I = (a, b)$ with derivative $g'(x)$ satisfying $|g'(x)| \leq L$ for $x \in I$, where we assume that $L < 1$. By Theorem 23.1 we know that $g(x)$ is Lipschitz continuous on I with Lipschitz constant L, and since $L < 1$, the function $g(x)$ has a unique fixed point $\bar{x} \in I$ satisfying $\bar{x} = g(\bar{x})$.

We know that $\bar{x} = \lim_{i\to\infty} x_i$, where $\{x_i\}_{i=1}^{\infty}$ is a sequence generated using Fixed Point Iteration: $x_{i+1} = g(x_i)$ for $i = 1, 2, ...$ To analyze the convergence of Fixed Point Iteration, we assume that $g(x)$ admits the following quadratic approximation close to $\bar{x}$ following the pattern of (24.26),

$$g(x) = g(\bar{x}) + g'(\bar{x})(x - \bar{x}) + \frac{1}{2}g''(\bar{x})(x - \bar{x})^2 + E_g(x, \bar{x}), \qquad (25.1)$$

where $|E_g(x, \bar{x})| \leq K_g(\bar{x})|x - \bar{x}|^3$. Choosing $x = x_i$, setting $e_i = x_i - \bar{x}$ and using $\bar{x} = g(\bar{x})$, we have for i large enough,

$$e_{i+1} = x_{i+1} - \bar{x} = g(x_i) - g(\bar{x}) = g'(\bar{x})e_i + \frac{1}{2}g''(\bar{x})e_i^2 + E_g(x_i, \bar{x}), \quad (25.2)$$

where $|E_g(x_i, \bar{x})| \leq K_g(\bar{x})|e_i|^3$. This formula gives an expansion of the error e_{i+1} at step $i + 1$ in terms of the different powers of e_i.

If $g'(\bar{x}) \neq 0$, then the linear term $g'(\bar{x})e_i$ dominates and

$$|e_{i+1}| \approx |g'(\bar{x})||e_i|, \qquad (25.3)$$

which says that the error decreases with (approximately) the factor $|g'(\bar{x})| < 1$ at each step, and we then say that the convergence is *linear*. If $g'(\bar{x}) = 0.1$, then we gain one decimal of accuracy in each step of Fixed Point Iteration.

As $|g'(\bar{x})|$ decreases, the convergence becomes faster. An extreme case arises when $g'(\bar{x}) = 0$. In this case, (25.2) implies

$$e_{i+1} = \frac{1}{2}g''(\bar{x})e_i^2 + E_g(x_i, \bar{x}),$$

so that neglecting the cubic term $E_g(x_i, \bar{x})$, we have

$$|e_{i+1}| \approx \frac{1}{2}|g''(\bar{x})|e_i^2. \qquad (25.4)$$

In this case the convergence is said to be *quadratic*, because the error $|e_{i+1}|$ is, up to the factor $|g''(\bar{x})/2|$, the square of the error $|e_i|$. If the convergence is quadratic, then the number of correct decimals roughly *doubles* in each step.

25.3 Newton's method

In Chapter Fixed Point Iteration, we saw that the problem of finding a root of an equation $f(x) = 0$, where $f(x)$ is a given function, can be reformulated as a fixed point equation $x = g(x)$, with $g(x) = x - \alpha f(x)$ and α a non-zero constant to choose. In fact, one may choose $\alpha(x)$ to depend in x and reformulate $f(x) = 0$ as

$$g(x) = x - \alpha(x)f(x),$$

if only $\alpha(\bar{x}) \neq 0$, where $\bar{x}$ is the root being computed. From above, we understand that a natural strategy is to choose α so as to make $g'(\bar{x})$ as small as possible. The ideal would be $g'(\bar{x}) = 0$. Differentiating the equation $g(x) = x - \alpha(x)f(x)$ with respect to x, we get

$$g'(x) = 1 - \alpha'(x)f(x) - \alpha(x)f'(x).$$

Assuming that $f'(\bar{x}) \neq 0$, and using $f(\bar{x}) = 0$,

$$\alpha(\bar{x}) = \frac{1}{f'(\bar{x})}.$$

Setting $\alpha(x) = \frac{1}{f'(x)}$ leads to *Newton's method* for computing a root of $f(x) = 0$: for $i = 0, 1, 2, ...$

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}, \tag{25.5}$$

where x_0 is a given initial root approximation. Newton's method corresponds to Fixed Point Iteration with

$$g(x) = x - \frac{f(x)}{f'(x)}. \tag{25.6}$$

Using Newton's method, it is natural to assume that $f'(\bar{x}) \neq 0$, which guarantees that $f'(x_i) \neq 0$ for i large if $f'(x)$ is Lipschitz continuous.

EXAMPLE 25.1. We apply Newton's method to compute the roots $\bar{x} = 2, 1, 0, -0.5, -1.5$ of the polynomial equation $f(x) = (x-2)(x-1)x(x+.5)(x+1.5) = 0$. We have that $f'(\bar{x}) \neq 0$ for all roots $\bar{x}$. We compute 21 Newton iterations for $f(x) = 0$ starting with 400 equally spaced initial values in $[-2, 3]$ then we mark the initial values x_0 with a color (shading) that indicates the root found by the iteration in Fig. 25.1. Each of the roots is contained in an interval in which all initial values produce convergence to the root. But outside these intervals the behavior of the iteration is unpredictable with near-by initial values converging to different roots.

25.4 Newton's method converges quadratically

We shall now prove that Newton's method converges quadratically if the initial approximation is good enough. We do this by computing the derivative of the corresponding fixed point function defined by (25.6):

$$g'(\bar{x}) = 1 - \frac{f'(\bar{x})^2 - f(\bar{x})f''(\bar{x})}{f'(\bar{x})^2} = \frac{f(\bar{x})f''(\bar{x})}{f'(\bar{x})^2} = 0,$$

where we used that $f(\bar{x}) = 0$ and the assumption that $f'(\bar{x}) \neq 0$. We conclude that Newton's method converges quadratically if $f'(\bar{x}) \neq 0$. This

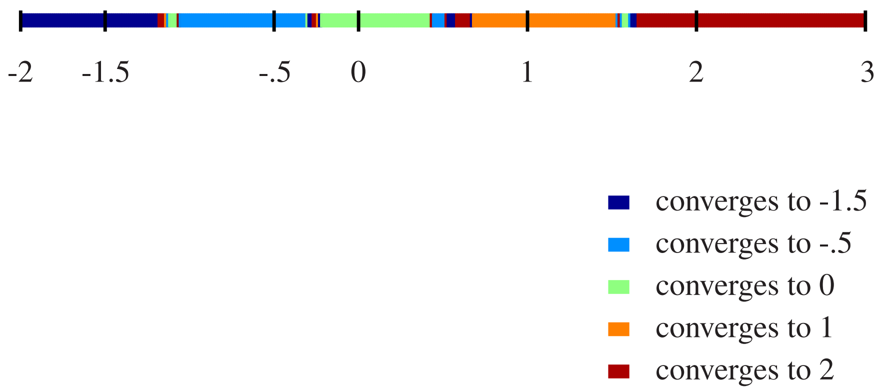


FIGURE 25.1. Each initial point for Newton's method applied to $f(x) = (x-2)(x-1)x(x+.5)(x+1.5)$ is marked with the indicated color (shade) depending on the result of the iteration.

result holds if we start sufficiently close to $\bar{x}$, so that in particular $f'(x_i) \neq 0$ for all i.

A more direct way to see that Newton's method converges quadratically, goes as follows. Subtract $\bar{x}$ from each side of (25.5) and use the fact that $f(x_i) = -f'(x_i)(\bar{x}-x_i) - E_f(\bar{x}, x_i)$, obtained from the linearization formula $f(\bar{x}) = f(x_i) + f'(x_i)(\bar{x} - x_i) + E_f(\bar{x}, x_i)$ because $f(\bar{x}) = 0$, to obtain

$$x_{i+1} - \bar{x} = x_i - \frac{f(x_i)}{f'(x_i)} - \bar{x} = \frac{E_f(\bar{x}, x_i)}{f'(x_i)}.$$

We conclude that

$$|x_{i+1} - \bar{x}| = |\frac{E_f(\bar{x}, x_i}{f'(x_i)}| \leq \frac{K_f}{|f'(x_i)|}|x_i - \bar{x}|^2,$$

which gives quadratic convergence if $f'(x)$ is bounded away from zero for x close to $\bar{x}$.

25.5 A geometric interpretation of Newton's method

There is an appealing geometric interpretation of Newton's method. Let x_i be an approximation of a root $\bar{x}$ of $f(x) = 0$ satisfying $f(\bar{x}) = 0$. Consider the tangent line to $y = f(x)$ at $x = x_i$,

$$y = f(x_i) + f'(x_i)(x - x_i).$$

Let x_{i+1} be the x-value where the tangent line crosses the x-axis, see Fig. 25.2, that is let x_{i+1} satisfy $f(x_i) + f'(x_i)(x_{i+1} - x_i) = 0$, so that

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}, \tag{25.7}$$

which is Newton's method. We conclude that the iterate x_{i+1} in Newton's method is the intersection of the tangent line to $f(x)$ at x_i with the x-axis. In words: trying to find $\bar{x}$, so that $f(\bar{x}) = 0$, we replace $f(x)$ by the linear approximation

$$\hat{f}(x) = f(x_i) + f'(x_i)(x - x_i),$$

that is by the tangent line at $x = x_i$, and then compute x_{i+1} as the solution of the equation $\hat{f}(x) = 0$. We shall find that this approach to Newton's method is easy to generalize to systems of equations corresponding to finding roots of $f(x)$ where $f : \mathbb{R}^n \to \mathbb{R}^n$.

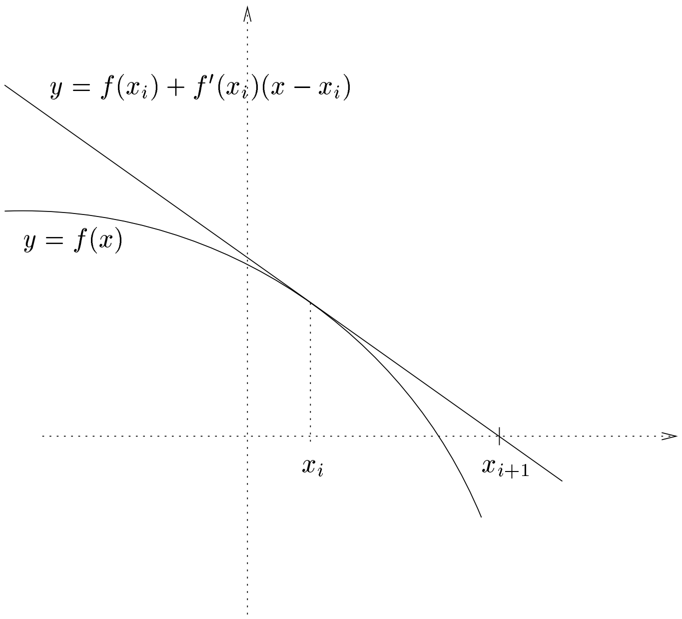


FIGURE 25.2. An illustration of one step of Newton's method from x_i to x_{i+1}.

25.6 What is the error in an approximate root?

Suppose x_i is an approximation of a root $\bar{x}$ of a given equation $f(x) = 0$. Can we say something about the error $x_i - \bar{x}$ from the knowledge of $f(x_i)$?

We will meet this question over and over again and we will refer to $f(x_i)$ as the *residual* of the approximation x_i. For the exact root $\bar{x}$, the residual is zero since $f(\bar{x}) = 0$, and for the approximation x_i, the residual $f(x_i)$ is not zero (unless by some miracle $x_i = \bar{x}$, or x_i is some root of $f(x) = 0$ different from $\bar{x}$).

Now, there is a very basic connection between the residual $f(x_i)$ and the error $x_i - \bar{x}$ that may be expressed as follows. Using the fact that $f(\bar{x}) = 0$ and assuming that $f(x)$ is differentiable at $\bar{x}$,

$$f(x_i) = f(x_i) - f(\bar{x}) = f'(\bar{x})(x_i - \bar{x}) + E_f(x_i, \bar{x}),$$

where $|E_f(x_i, \bar{x})| \leq K_f(\bar{x})|x_i - \bar{x}|^2$. Assuming that $f'(\bar{x}) \neq 0$, we conclude that

$$x_i - \bar{x} \approx \frac{f(x_i)}{f'(\bar{x})}, \tag{25.8}$$

up to the error term $(f'(\bar{x}))^{-1}E_f(x_i, \bar{x})$, which is quadratic in $x_i - \bar{x}$ and thus much smaller than $|x_i - \bar{x}|$ if x_i is close to $\bar{x}$, see Fig. 25.2.

The relation (25.8) shows that the root error $\bar{x}_i - \bar{x}$ is roughly proportional to the residual with the proportionality factor $(f'(\bar{x}))^{-1}$, if x_i is close to $\bar{x}$ and $f'(x)$ is Lipschitz continuous near $x = \bar{x}$. We summarize in the following basic theorem (the full proof of which will be given below using the Mean Value theorem).

Theorem 25.1 *If $f(x)$ is differentiable in an interval I containing a root $\bar{x}$ of $f(x) = 0$, and $|f'(x)|^{-1} \leq M$ for $x \in I$, then an approximate root $x_i \in I$, satisfies $|x_i - \bar{x}| \leq M|f(x_i)|$.*

In particular, if $f'(\bar{x})$ is very small, then the root error may be large although the residual is very small. In this case the process of computing the root $\bar{x}$ is said to be *ill-conditioned.*

EXAMPLE 25.2. We apply Newton's method to $f(x) = (x-1)^2 - 10^{-15}x$ with root $\bar{x} \approx 1.00000003162278$. Here $f'(1) = -10^{-15}$ and $f'(\bar{x}) \approx 0.0000000316$, so that $f'(x_n)$ is very small for all x_n close to $\bar{x}$, and the problem seems to be very ill-conditioned. We plot the errors and residuals versus iteration in Fig. 25.3. We see that the residuals become small quite a bit faster than the errors.

Introducing the approximation (25.8) into the definition of Newton's method,

$$x_{i+1} = x_i - f(x_i)/f'(x_i),$$

we get the relation

$$|x_i - \bar{x}| \approx |x_{i+1} - x_i|. \tag{25.9}$$

In other words, as an estimate of the error of $x_i - \bar{x}$, we can compute an extra step of Newton's method to get x_{i+1} and then use $|x_{i+1} - x_i|$ as an estimate of $|x_i - \bar{x}|$. This is an alternative way of estimating the root error $x_i - \bar{x}$, where the derivative $f'(x)$ does not enter explicitely.

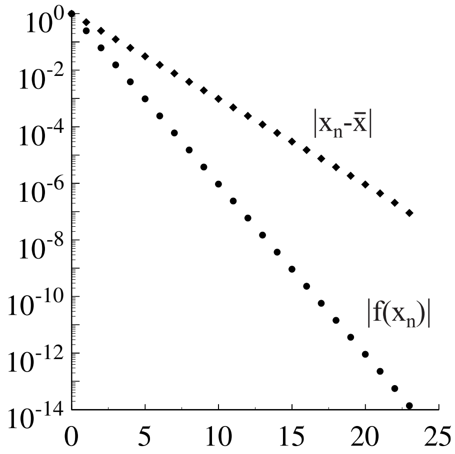


FIGURE 25.3. Plots of the residuals • and errors ⧫ versus iteration number for Newton's method applied to $f(x) = (x-1)^2 - 10^{-15}x$ with initial value $x_0 = 1$

EXAMPLE 25.3. We apply Newton's method to $f(x) = x^2 - 2$ and show the error and error estimate (25.9) in Fig. 25.4. The error estimate does a pretty good job.

i	$\lvert x_i - \bar{x}\rvert$	$\lvert x_{i+1} - x_i\rvert$
0	.586	.5
1	.086	.083
2	2.453×10^{-3}	2.451×10^{-3}
3	2.124×10^{-6}	2.124×10^{-6}
4	1.595×10^{-12}	1.595×10^{-12}
5	0	0

FIGURE 25.4. The error and error estimate for Newton's method for $f(x) = x^2 - 2$ with $x_0 = 2$.

25.7 Stopping criterion

Suppose we want to compute an approximation of a root $\bar{x}$ of a given equation $f(x) = 0$ with a certain accuracy, or *error tolerance* $TOL > 0$. In other words, suppose we want to guarantee that

$$|x_i - \bar{x}| \le TOL, \qquad (25.10)$$

where x_i is a computed approximation of the root $\bar{x}$. For example, we may choose $TOL = 10^{-m}$ corresponding to seeking an approximate root x_i with m correct decimals. Can we find some *stopping criterion* that tells us when to stop an iterative process with an approximation $\bar{x}_i$ satisfying (25.10)? The following criteria based on (25.8) presents itself: stop the iterative process at step i if

$$|(f'(\bar{x}_i))^{-1} f(\bar{x}_i)| \leq TOL. \tag{25.11}$$

Up to the change of argument from $\bar{x}$ to $\bar{x}_i$, this criterion guarantees the desired error control (25.10).

As an alternative stopping criterion for Newton's method, we may use (25.9), that is accept the approximation x_i with tolerance TOL if

$$|x_{i+1} - x_i| \leq TOL. \tag{25.12}$$

25.8 Globally convergent Newton methods

In this chapter, we have proved quadratic convergence of Newton's method under the assumption that we start close enough to the root of interest, that is we have prove *local convergence* of Newton's method. To get a sufficiently good initial approximation we may use the Bisection algorithm. Thus, by using the Bisection algorithm in an initial search of roots and then Newton's method for each individual root, we may obtain a *globally convergent* method combining efficiency (quadratic convergence) with reliability (guaranteed convergence).

Chapter 25 Problems

25.1. (a) Verify theoretically that the fixed point iteration for

$$g(x) = \frac{1}{2}\left(x + \frac{a}{x}\right)$$

with $\bar{x} = \sqrt{a}$ converges quadratically. (b) Try to say something about which initial values guarantee convergence for $a = 3$ by computing some fixed point iterations.

25.2. (a) Show analytically that Fixed Point Iteration for

$$g(x) = \frac{x(x^2 + 3a)}{3x^2 + a}$$

is third order convergent for computing $\bar{x} = \sqrt{a}$. (b) Compute a few iterations for $a = 2$ and $x_0 = 1$. How many digits of accuracy are gained with each iteration?

25.3. (a) Consider Newton's method applied to a differentiable function $f(x)$ with $f(\bar{x}) = f'(\bar{x}) = 0$, but $f''(\bar{x}) \neq 0$, that is $\bar{x}$ is a *double-root* of $f(x) = 0$. Prove that Newton's method in this case converges linearly, by proving that $g'(\bar{x}) = 1/2$, where $g(x) = x - f(x)/f'(x)$. (b) What is the rate of convergence of the following variant of Newton's method in the case of a double root: $g(x) = x - 2f(x)/f'(x)$? Hint: you may find it convenient to use l'Hopital's rule.

25.4. Use Newton's method to compute all the roots of $f(x) = x^5 + 3x^4 - 3x^3 - 5x^2 + 5x - 1$.

25.5. Use Newton's method to compute the smallest positive root of $f(x) = \cos(x) + \sin(x)^2(50x)$.

25.6. Use Newton's method to compute the root $\bar{x} = 0$ of the function

$$f(x) = \begin{cases} \sqrt{x} & x \geq 0 \\ -\sqrt{-x} & x < 0 \end{cases}$$

Does the method converge? If so, is it converging at second order? Explain your answer.

25.7. Apply Newton's method to $f(x) = x^3 - x$ starting with $x_0 = 1/\sqrt{5}$. Is the method converging? Explain your answer using a plot of $f(x)$.

25.8. (a) Derive an approximate relation between the residual $g(x) - x$ of a fixed point problem for g and the error of the fixed point iterate $x_n - \bar{x}$. (b) Devise two stopping criteria for a fixed point iteration. (c) Revise your fixed point code to make use of (a) and (b).

25.9. Use Newton's method to compute the root $\bar{x} = 1$ of $f(x) = x^4 - 3x^2 + 2x$. Is the method converging quadratically? Hint: you can test this by plotting $|x_n - 1|/|x_{n-1} - 1|$ for $n = 1, 2, \cdots$.

25.10. Assume that $f(x)$ has the form $f(x) = (x - \bar{x})^2 h(x)$ where h is a differentiable function with $h(\bar{x}) \neq 0$. (a) Verify that $f'(\bar{x}) = 0$ but $f''(\bar{x}) \neq 0$. (b) Show that Newton's method applied to $f(x)$ converges to $\bar{x}$ at a linear rate and compute the convergence factor.

26
Galileo, Newton, Hooke, Malthus and Fourier

> In a medium totally devoid of all resistance, all bodies would fall with the same speed. (Galileo)
>
> Everything that Galileo says about bodies falling in empty space is built without foundation; he ought first to have determined the nature of weight. (Descartes)
>
> When we have the decrees of Nature, authority goes for nothing. (Aristotle)
>
> Galileo has been the first to open the door to us to the whole realm of physics. (Hobbes)
>
> Provando e riprovando (Verify one and disprove the other). (Galileo)
>
> Measure what is measurable, and make measurable what is not so. (Galileo)

26.1 Introduction

In this chapter, we describe some basic models of physical phenomena that involve the derivatives of some function(s) and which therefore are called *differential equations*. The derivative is the fundamental tool for modeling in science and engineering. To formulate and solve differential equations has been a basic part of science since the days when Newton's Law of Motion was formulated. Today, the computer is opening new possibilities of modeling through differential equations, which Newton and all his scientific followers through the centuries, could never even dream of.

We thus formulate a couple of very basic differential models in this chapter, and we will spend a large part of the rest of this book to solve these equations or related generalizations using analytical or numerical techniques.

26.2 Newton's Law of motion

Newton's Law of motion is one of the cornerstones of the Newtonian physics that serves to describe much of the world that we live in. Newton's law states that the derivative with respect to time t of the *momentum* $m(t)v(t)$ of a body, which is the product of the *mass* $m(t)$ and the *velocity* $v(t)$ of the body, is equal to the *force* $f(t)$ acting on the body, that is,

$$(mv)'(t) = f(t). \tag{26.1}$$

If $m(t) = m$ is independent of time, then we can write Newton's Law as

$$mv'(t) = f(t), \tag{26.2}$$

or in its most familiar form:

$$ma(t) = f(t), \tag{26.3}$$

where

$$a(t) = v'(t),$$

is the *acceleration*.

Since the velocity is the derivative with respect to time of the *position* $s(t)$, that is, $v(t) = s'(t)$, we can write Newton's Law (26.2) in the case of constant mass, as follows:

$$ms''(t) = f(t). \tag{26.4}$$

If we think of the force $f(t)$ as a given function of time t, then (26.2) and (26.4) represent differential equations for the velocity $v(t)$ or the position $s(t)$. The differential equation thus involves some known function $(f(t))$, and the unknown is again a function ($v(t)$ or $s(t)$). The differential equation thus typically involves the derivative of an unknown function, and other given functions act as data. Note that typically we seek a function $s(t)$ satisfying the differential equation (26.4) not just at a particular instant of time t, but for all t in some interval.

26.3 Galileo's Law of motion

Galileo (1564-1642), mathematician, astronomer, philosopher, co-founder of the Scientific Revolution, performed his famous experiments dropping objects form the Tower of Pisa and counting the time for the objects to hit the ground, or using an inclining plane, see Fig. 26.1, to demonstrate his Laws of Motion. Galileo was condemned 1632 by the Catholic Church for questioning the idea that the Earth is the center of Universe, tried through these experiments to understand the nature of *motion*, a topic that already the Greeks were obsessed by.

FIGURE 26.1. Galileo demonstrating Laws of Motion

Galileo found that that an object close to the surface of the earth, that is acted upon only by the vertical gravity force, has a constant vertical acceleration independent of the mass or position of the body. This fact may be viewed as a special case of Newton's Law $ma(t) = f(t)$, where m is the mass of the body and $a(t)$ its vertical acceleration, and the gravity force $f(t)$ is given by $f(t) = mg$, where $g \approx 9.81$ (meter/sec^2), is the famous physical constant referred to as the *acceleration of gravity* at the surface of the Earth. Both Galileo and Newton would conclude that the acceleration $a(t) = g$ is independent of m and the position of the body as long as the body is not far from the surface of the Earth.

We now study the motion of one of the objects dropped vertically by Galileo from the Tower. Assuming that the mass of the object is m, and letting $s(t)$ denote the height above ground of the object at time t, with thus the positive direction upwards, see Fig. 26.2. In this coordinate system, a positive velocity $v(t) = s'(t)$ corresponds to the object moving upwards while a negative velocity means that the object is moving downwards. New-

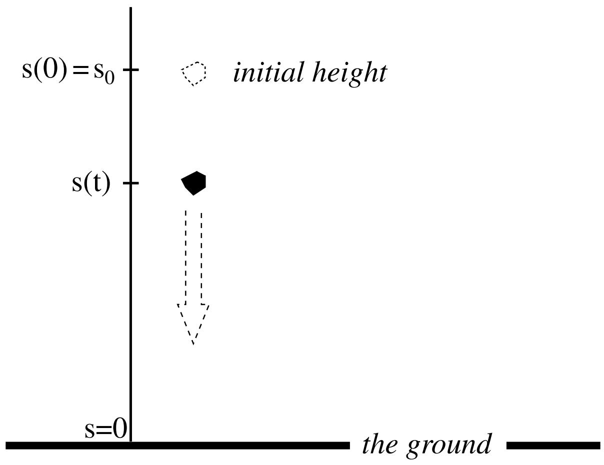


FIGURE 26.2. The coordinate system describing the position of a free falling object with the initial height $s(0) = s_0$ at time $t = 0$.

ton's Law states that

$$ms'' = -mg, \quad \text{or } mv'(t) = -mg$$

where a minus sign enters, because the gravity force is direct downwards. We thus have, dividing out the common factor of the mass m in the spirit of Galileo,

$$s''(t) = v'(t) = -g. \tag{26.5}$$

We conclude, using the fact $f'(t) = c$ for $f(t) = ct$ and c a constant, that

$$v(t) = -gt + c, \tag{26.6}$$

where c is a constant to determine. We check that $\frac{d}{dt}(-gt + c) = -g$ for all t. To pick out the one line that gives the velocity of the falling object in a concrete case, it is sufficient to know the velocity at some specific time. For example, if we suppose that the initial speed $v(0)$ at time $t = 0$ is known, $v(0) = v_0$, then we get the solution

$$v(t) = -gt + v_0, \quad \text{for } t > 0. \tag{26.7}$$

For example, if $v_0 = 0$, then the upward velocity is $v(t) = -gt$, and thus the downward velocity gt increases linearly with time. This is what Galileo observed (to his surprise).

Having now solved for the velocity $v(t)$ according to (26.7), we can now seek to find the position $s(t)$ by solving the differential equation $s'(t) = v(t)$, that is

$$s'(t) = -gt + v_0 \quad \text{for } t > 0. \tag{26.8}$$

Recalling that $(t^2)' = 2t$, and that $(cf(t))' = cf'(t)$, it is natural to believe that the solution of (26.8) would be the quadratic function

$$s(t) = -\frac{g}{2}t^2 + v_0 t + d,$$

where d is a constant, since this function satisfies (26.8). To determine $s(t)$ we need to determine the constant d. Typically, we can do this if we know $s(t)$ for some specific value of t, for example, if we know the initial position $s(0) = s_0$ at time $t = 0$. In this case we conclude that $d = s_0$, and thus we have the solution formulas

$$s(t) = -\frac{g}{2}t^2 + v_0 t + s_0, \quad v(t) = -gt + v_0 \quad \text{for } t > 0, \tag{26.9}$$

giving the position $s(t)$ and velocity $v(t)$ as functions of time t for $t > 0$, if the initial position $s(0) = s_0$ and velocity $v(0) = v_0$ are given. We summarize (the uniqueness of the solution will be settled below).

Theorem 26.1 *(*Galileo*) Let $s(t)$ and $v(t)$ denote the vertical position and vertical velocity of an object subject to free fall at time $t \geq 0$ with constant acceleration of gravity g, with the upward direction being positive. Then $s(t) = -\frac{g}{2}t^2 + v_0 t + s_0$ and $v(t) = -gt + v_0$, where $s(0) = s_0$ and $v(0) = v_0$ are given initial position and velocity.*

Example 26.1. If the initial height of the object is 15 meter and it is dropped from rest, what is the height at $t = .5$ sec? We have

$$s(.5) = -\frac{9.8}{2}(.5)^2 + 0 \times .5 + 15 = 13.775 \,\text{meter}$$

If initially it is thrown upwards at 2 meter/sec, the height at $t = .5$ sec is

$$s(.5) = -\frac{9.8}{2}(.5)^2 + 2 \times .5 + 15 = 14.775 \,\text{meter}$$

If initially it is thrown downwards at 2 meter/sec, the height at $t = .5$ sec is

$$s(.5) = -\frac{9.8}{2}(.5)^2 - 2 \times .5 + 15 = 12.775 \,\text{meter}$$

Example 26.2. An object starting from rest is dropped and hits the ground at $t = 5$ sec. What was its initial height? We have $s(5) = 0 = -\frac{9.8}{2}5^2 + 0 \times 5 + s_0$, and so $s_0 = 122.5$ meter.

26.4 Hooke's Law

Consider a spring of length L hanging vertically in equilibrium from the ceiling. Fix a coordinate system with x-axis directed downwards with the origin at the ceiling, and let $f(x)$ be the weight per unit length of the spring as a function of distance x to the ceiling. Imagine that the spring first is weightless (turn off gravity for a while) with corresponding zero tension force. Imagine then that you gradually turn on the gravity force and watch how the spring stretches under its increasing own weight. Let $u(x)$ be the corresponding *displacement* of a point in position x in the unstretched initial position. Can we determine the displacement $u(x)$ of the spring and the *tension force* $\sigma(x)$ in the spring as a function of x at full power of the gravity?

Hooke's Law for a linear (ideal) spring states that the tension force $\sigma(x)$ is proportional to the *deformation* $u'(x)$,

$$\sigma(x) = Eu'(x) \tag{26.10}$$

where $E > 0$ is a spring constant which we may refer to as the *modulus of elasticity* of the spring. Note that $u'(x)$ measures change of displacement $u(x)$ per unit of x, which is deformation.

The weight of the spring from position x to position $x + h$ for $h > 0$ is approximately equal to $f(x)h$, since $f(x)$ is weight per unit length, and this weight should be balanced by the difference of tension force:

$$-\sigma(x+h) + \sigma(x) \approx f(x)h, \quad \text{that is } -\frac{\sigma(x+h) - \sigma(x)}{h} \approx f(x),$$

which gives the *equilibrium equation* $-\sigma'(x) = f(x)$. Altogether, we get the following differential equation, since E is assumed constant independent of x,

$$-Eu''(x) = f(x), \quad \text{for } 0 < x < L. \tag{26.11}$$

Here $f(x)$ is a given function, and we seek the displacement $u(x)$ satisfying the differential equation (26.11). To determine $u(x)$ we need to specify that, for example, $u(0) = 0$, expressing that the spring is attached to the ceiling, and $u'(L) = 0$, expressing that the tension force is zero at the free end of the spring. If $f(x) = 1$ corresponding to a homogenous spring, and assuming $E = 1$ and $L = 1$, then we get the displacement $u(x) = x - \frac{x^2}{2}$, and the tension $\sigma(x) = 1 - x$.

Hooke (1635-1703) was A Curator at the Royal Society in London. Every week, except during the Summer vacation, Hooke had to demonstrate three or four experiments proving new laws of nature. Among other things, Hooke discovered the cellular structure of plants, the wave nature of light, Jupiter's red spots, and (probably before Newton!) the inverse square law of gravitation.

26.5 Newton's Law plus Hooke's Law

Consider a mass on a frictionless table connected to a spring attached to a wall. *Hooke's Law* for a spring states that the spring force arising when stretching or compressing the spring is proportional to the change of length of the spring from its rest state with zero force. We choose a horizontal x-axis in the direction of the spring so that the mass is located at $x = 0$ when the spring is at its rest state, and $x > 0$ corresponds to stretching the spring to the right, see Fig. 26.3. Let now $u(t)$ be the position of the

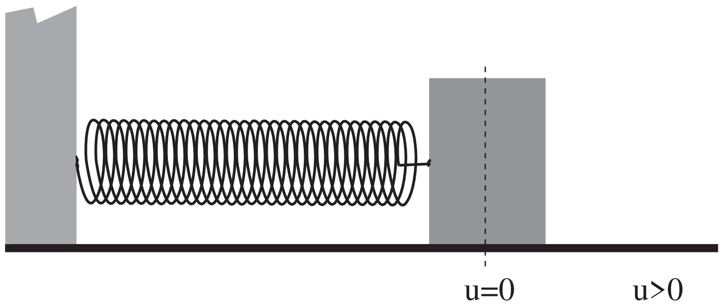


FIGURE 26.3. Illustration of the coordinate system used to describe a spring-mass system. The mass is allowed to slide freely back and forth with no friction.

mass at time $t > 0$. Hooke's Law states that the force $f(t)$ acting on the mass is given by

$$f(t) = -ku(t), \tag{26.12}$$

where the constant of proportionality $k = E/L > 0$ is the *spring constant*, with L the (natural) length of the spring. On the other hand, Newton's Law, states that $mu''(t) = f(t)$, and thus we get the following differential equation for the mass-spring system:

$$mu''(t) = -ku(t), \quad \text{or } mu''(t) + ku(t) = 0, \quad \text{for } t > 0. \tag{26.13}$$

We will see that solving this differential equation specifying the initial position $u(0)$ and the initial velocity $u'(0)$, will lead us into the world of the trigonometric functions $\sin(\omega x)$ and $\cos(\omega x)$ with $\omega = \sqrt{\frac{k}{m}}$.

26.6 Fourier's Law for heat flow

Fourier was one of the first mathematicians to study the process of *conduction of heat*. Fourier developed a mathematical technique for this purpose using *Fourier series*, where "general functions" are expressed as linear combinations of the *trigonometric functions* $\sin(nx)$ and $\cos(nx)$ with $n = 1, 2, ...$, see the Chapter Fourier series. Fourier created the simplest model of heat conduction stating that the *heat flow* is proportional to the *temperature difference*, or more generally *temperature gradient*, that is rate of change of temperature.

We now seek a simple model of the following situation: You live your life in Northern Scandinavia and need to electrically heat your car engine before being able to start the car in the morning. You may ask the question, when you should start the heating and what effect you should then turn on. Short time, high effect or long time small effect, depending on price of electric power which may vary over day and night.

A simple model could be as follows: let $u(t)$ be the temperature of the car engine at time t, let $q_+(t)$ be the flow of heat per unit time from the electrical heater to the engine, and let $q_-(t)$ be the heat loss from the engine to the environment (the garage). We then have

$$\lambda u'(t) = q_+(t) - q_-(t),$$

where $\lambda > 0$ is a constant which is referred to as the *heat capacity*. The heat capacity measures the rise of temperature of unit added heat.

Fourier's Law states that

$$q_-(t) = k(u(t) - u_0),$$

where u_0 is the temperature of the garage. Assuming that $u_0 = 0$ for simplicity, we thus get the following differential equation for the temperature $u(t)$:

$$\lambda u'(t) + ku(t) = q_+(t), \quad \text{for } t > 0, \ u(0) = 0, \tag{26.14}$$

where we added the initial condition $u(0) = 0$ stating that the engine has the temperature of the garage when the heater is turned on at time $t = 0$. We may now regard $q_+(t)$ as a given function, and λ and k as given functions, and seek the temperature $u(t)$ as a function of time.

26.7 Newton and rocket propulsion

Let's try to model the flight of a rocket far out in space away from any gravitational forces. When the rocket's engine is fired, the exhaust gases from the burnt fuel shoot backwards at high speed and the rocket moves forward so as to preserve the total momentum of the exhaust plus rocket, see Fig. 26.4. Assume the rocket moves to the right along a straight line which we identify with an x-axis. We let $s(t)$ denote the position of the rocket at time t and $v(t) = s'(t)$ its velocity and we let $m(t)$ denote the mass of the rocket at time t. We assume that the exhaust gases are ejected at a constant velocity $u > 0$ relative to the rocket in the direction of the negative x-axis. The total mass $m_e(t)$ of the exhaust at time t is $m_e(t) = m(0) - m(t)$, where $m(0)$ is the total mass of the rocket with fuel at initial time $t = 0$.

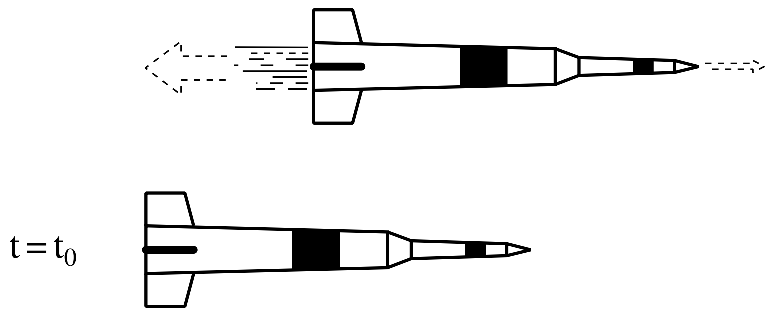


FIGURE 26.4. A model of rocket propulsion.

Let Δm_e be the exhaust released from the rocket during a time interval $(t, t+\Delta t)$. Conservation of total momentum of rocket plus exhaust over the time interval $(t, t+\Delta t)$, implies that

$$(m(t) - \Delta m_e)(v(t) + \Delta v) + \Delta m_e(v(t) - u) = m(t)v(t),$$

where Δv is the increase in speed of the rocket over $(t, t+\Delta t)$ and $\Delta m_e(v(t) - u)$ the momentum of the exhaust released during $(t, t+\Delta t)$. Recall that the momentum is the product of mass and velocity. This leads, using that $\Delta m_e = -\Delta m$ with Δm the change of rocket mass, to the relation

$$m(t)\Delta v = -u\Delta m.$$

Dividing by Δt, we are led to the differential equation

$$m(t)v'(t) = -um'(t) \quad \text{for } t > 0. \tag{26.15}$$

We shall show below that the solution can be expressed as (anticipating the use of the logarithm function $\log(x)$), assuming $v(0) = 0$,

$$v(t) = u \log\left(\frac{m(0)}{m(t)}\right), \tag{26.16}$$

connecting the velocity $v(t)$ to the mass $m(t)$. Typically, we may know $m(t)$ corresponding to firing the rocket in a specific way, and we may then determine the corresponding velocity $v(t)$ from (26.16). For example, we see that, when half the initial mass has been ejected at speed u relative to the rocket, then the speed of the rocket is equal to $u\log(2) \approx 0.6931\,u$. Note that this speed is less than u, because the speed of the exhaust ejected at time t varies from $-u$ for $t = 0$ to $v(t) - u$ for $t > 0$.

If the rocket engine would be fired so that the absolute speed of the exhaust was always $-u$ (that is at an increasing speed relative to the rocket), then conservation of momentum would state that

$$m(t)v(t) = (m(0) - m(t))u, \quad \text{that is } v(t) = u\left(\frac{m(0)}{m(t)} - 1\right).$$

In this case, the speed of the rocket would be equal to u when $m(t) = \frac{1}{2}m(0)$.

26.8 Malthus and population growth

Thomas Malthus (1766-1834), English priest and economist, developed a model for population growth in his treatise *An Essay on the Theory of Population*, Fig. 26.5. His study made him worried: the model indicated that the population would grow quicker than available resources. As a cure Malthus suggested late marriages. Malthus model was the following: Suppose the size of a population at time t is measured by a function $u(t)$, which is Lipschitz continuous in t. This means that we consider a quite large population and normalize so that the total number of individuals $P(t)$ of the population is set equal to $u(t)P_0$, where P_0 is the number of individuals at time $t = 0$, say, so that $u(0) = 1$. Malthus model for population growth is then

$$u'(t) = \lambda u(t), \quad \text{for } t > 0, \quad u(0) = 1, \tag{26.17}$$

where λ is a positive constant. Malthus thus assumes that the rate of growth $u'(t)$ is proportional to the population $u(t)$ at each given instant with a rate of growth factor λ, which may be the difference between the rates of birth and death (considering other factors such as migration negligible).

FIGURE 26.5. Thomas Malthus: "I think I may fairly make two postulata. First, That food is necessary to the existence of man. Secondly, That the passion between the sexes is necessary and will remain nearly in its present state".

Letting $t_n = kn$, $n = 0, 1, 2, ...$ be sequence of discrete time steps with constant time step $k > 0$, we can compare the differential equation (26.17), to the following discrete model

$$U_n = U_{n-1} + k\lambda U_{n-1}, \quad \text{for } n = 1, 2, 3..., \quad U_0 = 1. \tag{26.18}$$

Formally, (26.18) arises form $u(t_n) \approx u(t_{n-1}) + ku'(t_{n-1})$ inserting that $u'(t_{n-1}) = \lambda u(t_{n-1})$ and viewing U_n as an approximation of $u(t_n)$. We are familiar with the model (26.18), and we know that the solution is given by the formula

$$U_n = (1 + k\lambda)^n, \quad \text{for } n = 0, 1, 2, ... \tag{26.19}$$

We expect thus that

$$u(kn) \approx (1 + k\lambda)^n, \quad \text{for } n = 0, 1, 2, ... \tag{26.20}$$

or

$$u(t) \approx (1 + \frac{t}{n}\lambda)^n, \quad \text{for } n = 0, 1, 2, ... \tag{26.21}$$

We shall below prove that the differential equation (26.17) has a unique solution $u(t)$, which we will write as

$$u(t) = \exp(\lambda t) = e^{\lambda t}, \tag{26.22}$$

where $\exp(\lambda t) = e^{\lambda t}$ is the famous *exponential function*. We will prove that

$$\exp(\lambda t) = e^{\lambda t} = \lim_{n\to\infty}(1 + \frac{\lambda t}{n})^n, \tag{26.23}$$

corresponding to the fact that U_n will be an increasingly good approximation of $u(t)$ for a given $t = nk$, as n tends infinity and thus the time step $k = \frac{t}{n}$ tends to zero.

In particular, we have setting $\lambda = 1$, that $u(t) = \exp(t)$ is the solution of the differential equation

$$u'(t) = u(t), \quad \text{for } t > 0, \quad u(0) = 1. \tag{26.24}$$

We plotted the function $\exp(t)$ in Fig. 4.4, from which the worries of Malthus become understandable. The exponential function grows really quickly after some time!

Many physical situations are modeled by (26.17) with $a \in \mathbb{R}$, such as earnings from compound interest ($a > 0$), and radioactive decay ($a < 0$).

26.9 Einstein's Law of motion

Einstein's version of Newton's Law $(m_0 v(t))'(t) = f(t)$, modeling the motion of a particle of mass m_0 moving in a straight line with velocity $v(t)$ under the influence of a force $f(t)$, is

$$\frac{d}{dt}\left(\frac{m_0}{\sqrt{1 - v^2(t)/c^2}}v(t)\right) = f(t), \tag{26.25}$$

where m_0 is the mass of the particle at zero speed and $c \approx 3 \times 10^8$ m/s is the speed of light in a vacuum. Setting

$$w(t) = \frac{v(t}{\sqrt{1 - v^2(t)/c^2}},$$

we can write Einstein's equation

$$w'(t) = \frac{f(t)}{m_0},$$

and assuming $f(t) = F$ is constant, and $v(0) = 0$, we obtain $w(t) = \frac{F}{m_0}t$, which gives the following equation for $v(t)$

$$\frac{v(t)}{\sqrt{1 - v^2(t)/c^2}} = \frac{F}{m_0}t.$$

Squaring both sides, we get

$$\frac{v^2(t)}{1 - v^2(t)/c^2} = \frac{F^2}{m_0^2}t^2.$$

and solving for $v(t)$, we get

$$v^2(t) = c^2 \, \frac{F^2 t^2}{m_0^2 c^2 + F^2 t^2}. \tag{26.26}$$

We conclude that $v^2(t) < c^2$ for all t: Einstein says that no object can be accelerated to a velocity bigger or equal to the velocity of light!

To determine the position of the particle, we therefore have to solve the differential equation

$$s'(t) = v(t) = \frac{cFt}{\sqrt{m_0^2 c^2 + F^2 t^2}}. \tag{26.27}$$

We will return to this task below.

26.10 Summary

We have above derived differential equation models of the following basic forms:

$$u'(x) = f(x) \quad \text{for } x > 0,\ u(0) = u_0, \tag{26.28}$$

$$u'(x) = u(x) \quad \text{for } x > 0,\ u(0) = u_0, \tag{26.29}$$

$$u''(x) + u(x) = 0 \quad \text{for } x > 0,\ u(0) = u_0,\ u'(0) = u_1, \tag{26.30}$$

$$-u''(x) = f(x) \quad \text{for } 0 < x < 1,\ u(0) = 0,\ u'(1) = 0. \tag{26.31}$$

Here we view the function $f(x)$, and the constants u_0 and u_1 as given, and we seek the unknown function $u(x)$ satisfying the differential equation for $x > 0$.

We shall see in Chapter *The Fundamental Theorem of Calculus* that the solution $u(x)$ to (26.28) is given by

$$u(x) = \int_0^x f(y)\,dy + u_0.$$

Integrating twice, we can the write the solution $u(x)$ of (26.31) in terms of the data $f(x)$ as

$$u(x) = \int_0^x g(y)\,dy,\; u'(y) = g(y) = -\int_1^y f(z)\,dz, \text{ for } 0 < x < 1.$$

We shall see in Chapter The exponential function $\exp(x)$ that the solution to (26.29) is $u(x) = u_0 \exp(x)$. We shall see in Chapter *Trigonometric functions* that the solution $u(x)$ to (26.30) is $\sin(x)$ if $u_0 = 0$ and $u_1 = 1$, and $\cos(x)$ if $u_0 = 1$ and $u_1 = 0$.

We conclude that the basic elementary functions $\exp(x)$, $\sin(x)$ and $\cos(x)$ are solutions to basic differential equations. We shall see that it is fruitful to define these elementary functions as the solutions of the corresponding differential equations, and that the basic properties of the the functions can easily be derived using the differential equations. For example, the fact that $D\exp(x) = \exp(x)$ follows from the defining differential equation $Du(x) = u(x)$. Further, the fact that $D\sin(x) = \cos(x)$ follows by differentiating $D^2\sin(x) + \sin(x) = 0$ with respect to x to get $D^2(D\sin(x)) + D\sin(x) = 0$, which expresses that $D\sin(x)$ solves $u''(x) + u(x) = 0$ with initial conditions $u_0 = 1$ and $u_1 = Du(0) = D^2\sin(0) = -\sin(0) = 0$, that is $D\sin(x) = \cos(x)$!

Chapter 26 Problems

26.1. An object is dropped from height $s_0 = 15$ m. After how long time does it hit the ground? How much does an initial velocity $v_0 = -2$ delay the time of hitting the ground? Use the (crude) approximation $g \approx 10$ to simplify the analysis. Which initial velocity v_0 is required to double the time of the fall?

26.2. Find the displacement of a hanging spring of length $L = 1$ and $E = 1$ with a non-uniform weight given by $f(x) = x$.

26.3. (Tricky!) Find the displacement of a hanging spring of length $L = 2$ which have been tied together of two springs of length $L = 1$ with (a) $E = 1$ and $E = 2$, respectively, assuming the weight of the springs to be uniform with $f(x) = 1$, (b) both with $E = 1$ but with weights $f = 1$ and $f = 2$ per unit length, respectively.

26.4. Derive (26.12) from (26.10). Hint: u as in (26.12) corresponds to the displacement in (26.10) at $x = L$ being $L + u$, and the displacement being linear in x.

26.5. Develop more complicated population models in the form of systems of differential equations.

26.6. Derive differential equation models for systems of springs coupled in series or in parallel.

26.7. Derive a differential equation model for a "bungee jump" of a body hooked up to a fixed support with a rubber band and attracted by gravity.

26.8. You leave a hot cup of coffee on the table in a room with temperature 20^o Celsius. Put up an equation for the rate of change of the temperature of the coffee with time.

Part I

Integrals
Geometry in $\mathbb{R}^n$

$$u(x_N) - u(x_0) = \int_{x_0}^{x_N} u'(x)\,dx$$
$$\approx \sum_{j=1}^{N} u'(x_{j-1})(x_j - x_{j-1})$$

$$a \cdot b = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

27
The Integral

The two questions, the first that of finding the description of the curve from its elements, the second that of finding the figure from the given differences, both reduce to the same thing. From this it can be taken that the whole of the theory of the inverse method of the tangents is reducible to quadratures. (Leibniz 1673)

Utile erit scribit $\int$ pro omnia. (Leibniz, October 29 1675)

27.1 Primitive functions and integrals

In this chapter, we begin the study of the subject of *differential equations*, which is one of the common ties binding together all areas of science and engineering, and it would be hard to overstate its importance. We have been preparing for this chapter for a long time, starting from the beginning with Chapter *A very short course in Calculus*, through all of the chapters on functions, sequences, limits, real numbers, derivatives and basic differential equation models. So we hope the gentle reader is both excited and ready to embark on this new exploration.

We begin our study with the simplest kind of differential equation, which is of fundamental importance:

> Given the **function** $f : I \to \mathbb{R}$ defined on the interval $I = [a, b]$, find a **function** $u(x)$ on I, such that the derivative $u'(x)$ of $u(x)$ is equal to $f(x)$ for $x \in I$.

We can formulate this problem more concisely as: given $f : I \to \mathbb{R}$ find $u : I \to \mathbb{R}$ such that

$$u'(x) = f(x) \tag{27.1}$$

for all $x \in I$. We call the solution $u(x)$ of the differential equation $u'(x) = f(x)$ for $x \in I$, a *primitive function* of $f(x)$, or an *integral* of $f(x)$. Sometimes the term *antiderivative* is also used.

To understand what we mean by "solving" (27.1), we consider two simple examples. If $f(x)$=1 for $x \in \mathbb{R}$, then $u(x) = x$ is a solution of $u'(x) = f(x)$ for $x \in \mathbb{R}$, since $Dx = 1$ for all $x \in \mathbb{R}$. Likewise if $f(x) = x$, then $u(x) = x^2/2$ is a solution of $u'(x) = f(x)$ for $x \in \mathbb{R}$, since $Dx^2/2 = x$ for $x \in \mathbb{R}$. Thus the function x is a primitive function of the constant function 1, and $x^2/2$ is a primitive function of the function x.

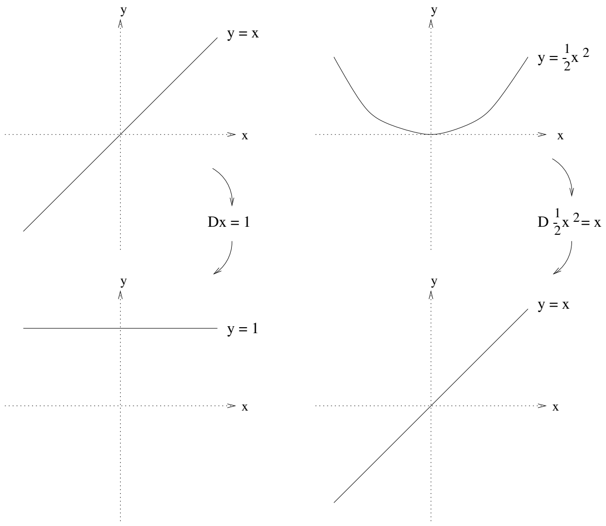


FIGURE 27.1. $Dx = 1$ and $D(x^2/2) = x$.

We emphasize that the solution of (27.1) is a **function** defined on an interval. We can interpret the problem in physical terms if we suppose that $u(x)$ represents some accumulated quantity like a sum of money in a bank, or an amount of rain, or the height of a tree, while x represents some changing quantity like time. Then solving (27.1) amounts to computing the total accumulated quantity $u(x)$ from knowledge of the rate of growth $u'(x) = f(x)$ at each instant x. This interpretation suggests that finding the total accumulated quantity $u(x)$ amounts to adding little pieces of momentary increments or changes of the quantity $u(x)$. Thus we expect that finding the integral $u(x)$ of a function $f(x)$ satisfying $u'(x) = f(x)$ will amount to some kind of *summation*.

A familiar example of this problem occurs when $f(x)$ is a velocity and x represents time so that the solution $u(x)$ of $u'(x) = f(x)$, represents the distance traveled by a body moving with instantaneous velocity $u'(x) = f(x)$. As the examples above show, we can solve this problem in simple cases, for example when the velocity $f(x)$ is equal to a constant v for all x and therefore the distance traveled during a time x is $u(x) = vx$. If we travel with constant velocity 4 miles/hour for two hours, then the distance traveled is 8 miles. We reach these 8 miles by accumulating distance foot-by-foot, which would be very apparent if we are walking!

An important observation is that the differential equation (27.1) alone is not sufficient information to determine the solution $u(x)$. Consider the interpretation when f represents velocity and u distance traveled by a body. If we want to know the position of the body, we need to know only the distance traveled but also the starting position. In general, a solution $u(x)$ to (27.1) is determined only up to a constant, because the derivative of a constant is zero. If $u'(x) = f(x)$, then also $(u(x) + c)' = f(x)$ for any constant c. For example, both $u(x) = x^2$ and $u(x) = x^2 + 1$ satisfy $u'(x) = 2x$. Graphically, we can see that there are many "parallel" functions that have the same slope at every point. The constant may be specified by

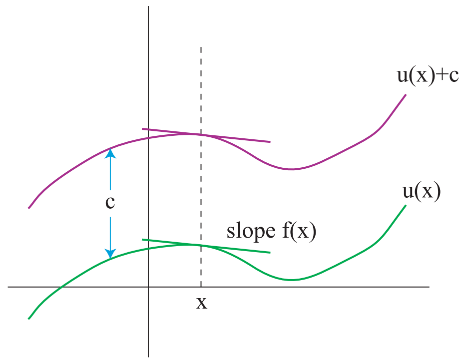


FIGURE 27.2. Two functions that have the same slope at every point.

specifying the value of the function $u(x)$ at some point. For example, the solution of $u'(x) = x$ is $u(x) = x^2 + c$ with c a constant, and specifying $u(0) = 1$ gives that $c = 1$.

More generally, we now formulate our basic problem as follows: Given $f : [a, b] \to \mathbb{R}$ and u_a, find $u : [a, b] \to \mathbb{R}$ such that

$$\begin{cases} u'(x) = f(x) & \text{for } a < x \leq b, \\ u(a) = u_a, \end{cases} \tag{27.2}$$

where u_a is a given *initial value*. The problem (27.2) is the simplest example of an *initial value problem* involving a differential equation and an initial value. The terminology naturally couples to situations in which x represents time and $u(a) = u_a$ amounts to specifying $u(x)$ at the initial time $x = a$. Note that we often keep the initial value terminology even if x represents a quantity different from time, and in case x represents a space coordinate we may alternatively refer to (27.2) as a *boundary value problem* with now $u(a) = u_a$ representing a given *boundary value*

We shall now prove that the initial value problem (27.2) has a unique solution $u(x)$ if the given function $f(x)$ is Lipschitz continuous on $[a, b]$. This is the *Fundamental Theorem of Calculus*, which stated in words says that a Lipschitz continuous function has a (unique) primitive function. Leibniz referred to the Fundamental Theorem as the "inverse method of tangents" because he thought of the problem as trying to find a curve $y = u(x)$ given the slope $u'(x)$ of its tangent at every point x.

We shall give a constructive proof of the Fundamental Theorem, which not only proves that $u : I \to \mathbb{R}$ exists, but also gives a way to compute $u(x)$ for any given $x \in [a, b]$ to any desired accuracy by computing a sum involving values of $f(x)$. Thus the version of the Fundamental Theorem we prove contains two results: (i) the existence of a primitive function and (ii) a way to compute a primitive function. Of course, (i) is really a consequence of (ii) since if we know how to compute a primitive function, we also know that it exists. These results are analogous to defining $\sqrt{2}$ by constructing a Cauchy sequence of approximate solutions of the equation $x^2 = 2$ by the Bisection algorithm. In the proof of the Fundamental Theorem we shall also construct a Cauchy sequence of approximate solutions of the differential equation (27.2) and show that the limit of the sequence is an exact solution of (27.2).

We shall express the solution $u(x)$ of (27.2) given by the Fundamental Theorem in terms of the data $f(x)$ and u_a as follows:

$$u(x) = \int_a^x f(y)\, dy + u_a \quad \text{for } a \leq x \leq b, \tag{27.3}$$

where we refer to

$$\int_a^x f(y)\, dy$$

as the *integral* of f over the interval $[a, x]$, a and x as the *lower and upper limits of integration* respectively, $f(y)$ as the *integrand* and y the *integration variable*. This notation was introduced on October 29 1675 by Leibniz, who thought of the integral sign $\int$ as representing "summation" and dy as the "increment" in the variable y. The notation of Leibniz is part of the big success of Calculus in science and education, and (like a good cover of a record) it gives a direct visual expression of the mathematical content of the integral in very suggestive form that indicates both the construction of the integral and how to operate with integrals. Leibniz choice of notation plays an important role in making Calculus into a "machine" which "works by itself".

We recapitulate: There are two basic problems in Calculus. The first problem is to determine the derivative $u'(x)$ of a given function $u(x)$. We have met this problem above and we know a set of rules that we can use to attack this problem. The other problem is to find a function $u(x)$ given its derivative $u'(x)$. In the first problem we assume knowledge of $u(x)$ and we want to find $u'(x)$. In the second problem we assume knowledge of $u'(x)$ and we want to find $u(x)$.

As an interesting aside, the proof of the Fundamental Theorem also shows that the integral of a function over an interval may be interpreted as the area underneath the graph of the function over the interval. This couples the problem of finding a primitive function, or computing an integral, to that of computing an area, that is to *quadrature*. We expand on this geometric interpretation below.

Note that in (27.2), we require the differential equation $u'(x) = f(x)$ to be satisfied for x in the half-open interval $(a, b]$ excluding the left end-point $x = a$, where the differential equation is replaced by the specification $u(a) = u_a$. The proper motivation for this will become clear as we develop the proof of the Fundamental Theorem. Of course, the derivative $u'(b)$ at the right end-point $x = b$, is taken to be the left-hand derivative of u. By continuity, we will in fact have also $u'(a) = f(a)$, with $u'(a)$ the right-hand derivative.

27.2 Primitive function of $f(x) = x^m$ for $m = 0, 1, 2, ...$

For some special functions $f(x)$, we can immediately find primitive functions $u(x)$ satisfying $u'(x) = f(x)$ for x in some interval. For example, if $f(x) = 1$, then $u(x) = x + c$, with c a constant, for $x \in \mathbb{R}$. Further, if $f(x) = x$, then $u(x) = x^2/2 + c$ for $x \in \mathbb{R}$. More generally, if $f(x) = x^m$, where $m = 0, 1, 2, 3, ..$, then $u(x) = x^{m+1}/(m + 1) + c$. Using the notation (27.3) for $x \in \mathbb{R}$ we write

$$\int_0^x 1\,dy = x, \quad \int_0^x y\,dy = \frac{x^2}{2}, \tag{27.4}$$

and more generally for $m = 0, 1, 2, ...,$

$$\int_0^x y^m\,dy = \frac{x^{m+1}}{m+1}, \tag{27.5}$$

because both right and left hand sides vanish for $x = 0$.

27.3 Primitive function of $f(x) = x^m$ for $m = -2, -3, \ldots$

We recall that if $v(x) = x^{-n}$, where $n = 1, 2, 3, ..$, then $v'(x) = -nx^{-(n+1)}$, where now $x \neq 0$. Thus a primitive function of $f(x) = x^m$ for $m = -2, -3, ..$ is given by $u(x) = x^{m+1}/(m+1)$ for $x > 0$. We can state this fact as follows: For $m = -2, -3, ...,$

$$\int_1^x y^m \, dy = \frac{x^{m+1}}{m+1} - \frac{1}{m+1} \quad \text{for } x > 1, \tag{27.6}$$

where we start the integration arbitrarily at $x = 1$. The starting point really does not matter as long as we avoid 0. We have to avoid 0 because the function x^m with $m = -2, -3, ..,$, tends to infinity as x tends to zero. To compensate for starting at $x = 1$, we subtract the corresponding value of $x^{m+1}/(m+1)$ at $x = 1$ from the right hand side. We can write analogous formulas for $0 < x < 1$ and $x < 0$.

Summing up, we see that the polynomials x^m with $m = 0, 1, 2, ...$, have the primitive functions $x^{m+1}/(m+1)$, which again are polynomials. Further, the rational functions x^m for $m = -2, -3, ...$, have the primitive functions $x^{m+1}/(m+1)$, which again are rational functions.

27.4 Primitive function of $f(x) = x^r$ for $r \neq -1$ and $x > 0$

Given our success so far, it would be easy to get overconfident. But we encounter a serious difficulty even with these early examples. Extending the previous arguments to rational powers of x, since $Dx^s = sx^{s-1}$ for $s \neq 0$ and $x > 0$, we have for $r = s - 1 \neq -1$,

$$\int_1^x y^r \, dy = \frac{x^{r+1}}{r+1} - \frac{1}{r+1} \quad \text{for } x > 1. \tag{27.7}$$

This formula breaks down for $r = -1$ and therefore we do not know a primitive function of $f(x) = x^r$ with $r = -1$ and moreover we don't even know that one exists. In fact, it turns out that most of the time we cannot solve the differential equation (27.2) in the sense of writing out $u(x)$ in terms of known functions. Being able to integrate simple rational functions is special. The Fundamental Theorem of Calculus will give us a way past this difficulty by providing the means to approximate the unknown solution to any desired accuracy.

27.5 A quick overview of the progress so far

Any function obtained by linear combinations, products, quotients and compositions of functions of the form x^r with rational power $r \neq 0$ and $x > 0$, can be differentiated analytically. If $u(x)$ is such a function, we thus obtain an analytical formula for $u'(x)$. If we now choose $f(x) = u'(x)$, then of course $u(x)$ satisfies the differential equation $u'(x) = f(x)$, so that we can write recalling Leibniz notation:

$$u(x) = \int_0^x f(y)\,dy + u(0) \quad \text{for } x \geq 0,$$

which apparently states that the function $u(x)$ is a primitive function of its derivative $f(x) = u'(x)$ (assuming that $u(x)$ is defined for all $x \geq 0$ so that no denominator vanishes for $x \geq 0$).

We give an example: Since $D(1+x^3)^{\frac{1}{3}} = (1+x^3)^{-\frac{2}{3}}x^2$ for $x \in \mathbb{R}$, we can write

$$(1+x^3)^{\frac{1}{3}} = \int_0^x \frac{y^2}{(1+y^3)^{\frac{2}{3}}}\,dy + 1 \quad \text{for } x \in \mathbb{R}.$$

In other words, we know primitive functions $u(x)$ satisfying the differential equation $u'(x) = f(x)$ for $x \in I$, for any function $f(x)$, which itself is a derivative of some function $v(x)$ so that $f(x) = v'(x)$ for $x \in I$. The relation between $u(x)$ and $v(x)$ is then

$$u(x) = v(x) + c \quad \text{for } x \in I,$$

for some constant c.

On the other hand, if $f(x)$ is an arbitrary function of another from, then we may not be able to produce an analytical formula for the corresponding primitive function $u(x)$ very easily or not at all. The Fundamental Theorem now tells us how to compute a primitive function of an arbitrary Lipschitz continuous function $f(x)$. We shall see that in particular, the function $f(x) = x^{-1}$ has a primitive function for $x > 0$ which is the famous *logarithm function* $\log(x)$. The Fundamental Theorem therefore gives in particular a constructive procedure for computing $\log(x)$ for $x > 0$.

27.6 A “very quick proof” of the Fundamental Theorem

We shall now enter into the proof of the Fundamental Theorem. It is a good idea at this point to review the Chapter *A very short course in Calculus*. We shall give a sequence of successively more complete versions of the proof of the Fundamental Theorem with increasing precision and generality in each step.

The problem we are setting out to solve has the following form: given a function $f(x)$, find a function $u(x)$ such that $u'(x) = f(x)$ for all x in an interval. In this problem, we start with $f(x)$ and seek a function $u(x)$ such that $u'(x) = f(x)$. However in the early "quick" versions of the proofs, it will appear that we have turned the problem around by starting with a given function $u(x)$, differentiating u to get $f(x) = u'(x)$, and then recovering $u(x)$ as a primitive function of $f(x) = u'(x)$. This naturally appears to be quite meaningless circular reasoning, and some Calculus books completely fall into this trap. But we are doing this to make some points clear. In the final proof, we will in fact start with $f(x)$ and construct a function $u(x)$ that satisfies $u'(x) = f(x)$ as desired!

Let now $u(x)$ be differentiable on $[a, b]$, let $x \in [a, b]$, and let $a = y_0 < y_1 < ... < y_m = x$ be a *subdivision* of $[a, x]$ into subintervals $[a, y_1)$, $[y_1, y_2)$,...,$[y_{m-1}, x)$. By repeatedly subtracting and adding $u(y_j)$, we obtain the following identity which we refer to as a *telescoping sum* with the terms cancelling two by two:

$$\begin{aligned} u(x) - u(a) &= u(y_m) - u(y_0) \\ &= u(y_m) - u(y_{m-1}) + u(y_{m-1}) - u(y_{m-2}) + u(y_{m-2}) \\ &\qquad - \cdots + u(y_2) - u(y_1) + u(y_1) - u(y_0). \end{aligned} \tag{27.8}$$

We can write this identity in the form

$$u(x) - u(a) = \sum_{i=1}^{m} \frac{u(y_i) - u(y_{i-1})}{y_i - y_{i-1}} (y_i - y_{i-1}), \tag{27.9}$$

or as

$$u(x) - u(a) = \sum_{i=1}^{m} f(y_{i-1})(y_i - y_{i-1}), \tag{27.10}$$

if we set

$$f(y_{i-1}) = \frac{u(y_i) - u(y_{i-1})}{y_i - y_{i-1}} \quad \text{for } i = 1, ..., m. \tag{27.11}$$

Recalling the interpretation of the derivative as the ratio of the change in a function to a change in its input, we obtain our first version of the Fundamental Theorem as the following analog of (27.10) and (27.11):

$$u(x) - u(a) = \int_a^x f(y)\, dy \quad \text{where} \quad f(y) = u'(y) \quad \text{for } a < y < x.$$

In the integral notation, the sum $\sum$ corresponds to the integral sign $\int$, the increments $y_i - y_{i-1}$ correspond to dy, the y_{i-1} to the integration variable y, and the difference quotient $\frac{u(y_i) - u(y_{i-1})}{y_i - y_{i-1}}$ corresponds to the derivative $u'(y_{i-1})$.

This is the way that Leibniz was first led to the Fundamental Theorem at the age of 20 (without having studied any Calculus at all) as presented in his *Art of Combinations* from 1666.

Note that (27.8) expresses the idea that “the whole is equal to the sum of the parts” with “the whole” being equal to $u(x)-u(a)$ and the “parts” being the differences $(u(y_m)-u(y_{m-1}))$, $(u(y_{m-1})-u(y_{m-2}))$,...,$(u(y_2)-u(y_1))$ and $(u(y_1)-u(y_0))$. Compare to the discussion in Chapter *A very short Calculus course* including Leibniz' teen-age dream.

27.7 A “quick proof” of the Fundamental Theorem

We now present a more precise version of the above “proof”. To exercise flexibility in the notation, which is a useful ability, we change notation slightly. Let $u(x)$ be uniformly differentiable on $[a,b]$, let $\bar{x}\in[a,b]$, and let $a=x_0<x_1<...<x_m=\bar{x}$ be a partition of $[a,\bar{x}]$. We thus change from y to x and from x to $\bar{x}$. With this notation x serves the role of a variable and $\bar{x}$ is a particular value of x. We recall the identity (27.9) in its new dress:

$$u(\bar{x})-u(a)=\sum_{i=1}^{m}\frac{u(x_i)-u(x_{i-1})}{x_i-x_{i-1}}(x_i-x_{i-1}). \tag{27.12}$$

By the uniform differentiability of u:

$$u(x_i)-u(x_{i-1})=u'(x_{i-1})(x_i-x_{i-1})+E_u(x_i,x_{i-1}),$$

where

$$|E_u(x_i,x_{i-1})|\leq K_u(x_i-x_{i-1})^2, \tag{27.13}$$

with K_u a constant, we can write the identity as follows:

$$u(\bar{x})-u(a)=\sum_{i=1}^{m}u'(x_{i-1})(x_i-x_{i-1})+\sum_{i=1}^{m}E_u(x_i,x_{i-1}). \tag{27.14}$$

Setting h equal to the largest increment x_i-x_{i-1}, so that $x_i-x_{i-1}\leq h$ for all i, we find

$$\sum_{i=1}^{m}|E_u(x_i,x_{i-1})|\leq\sum_{i=1}^{m}K_u(x_i-x_{i-1})h=K_u(\bar{x}-a)h.$$

The formula (27.14) can thus be written

$$u(\bar{x})-u(a)=\sum_{i=1}^{m}u'(x_{i-1})(x_i-x_{i-1})+E_h, \tag{27.15}$$

where

$$|E_h|\leq K_u(\bar{x}-a)h. \tag{27.16}$$

The Fundamental Theorem is the following analog of this formula:

$$u(\bar{x}) - u(a) = \int_a^{\bar{x}} u'(x)\, dx, \qquad (27.17)$$

with the sum $\sum$ corresponding to the integral sign $\int$, the increments $x_i - x_{i-1}$ corresponding to dx, and x_i corresponding to the integration variable x. We see by (27.16) that the additional term E_h in (27.15) tends to zero as the maximal increment h tends to zero. We thus expect (27.17) to be a limit form of (27.15) as h tends to zero.

27.8 A proof of the Fundamental Theorem of Calculus

We now give a full proof of the Fundamental theorem. We assume for simplicity that $[a, b] = [0, 1]$ and the initial value $u(0) = 0$. We comment on the general problem at the end of the proof. So the problem we consider is: Given a Lipschitz continuous function $f : [0, 1] \to \mathbb{R}$, find a solution $u(x)$ of the initial value problem,

$$\begin{cases} u'(x) = f(x) & \text{for } 0 < x \le 1, \\ u(0) = 0. \end{cases} \qquad (27.18)$$

We shall now construct an approximation to the solution $u(x)$ and thereby give a meaning to the solution formula

$$u(\bar{x}) = \int_0^{\bar{x}} f(x)\, dx \quad \text{for } 0 \le \bar{x} \le 1.$$

To this end, let n be a natural number and let $0 = x_0 < x_1 < ... < x_N = 1$ be the subdivision of the interval $[0, 1]$ with *nodes* $x_i^n = ih_n$, $i = 0, ..., N$, where $h_n = 2^{-n}$ and $N = 2^n$. We thus divide the given interval $[0, 1]$ into subintervals $I_i^n = (x_{i-1}^n, x_i^n]$ of equal lengths $h_n = 2^{-n}$, see Fig. 27.3.

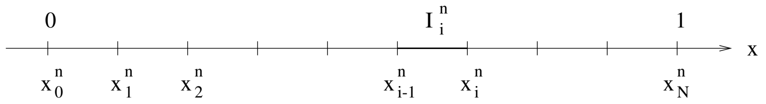


FIGURE 27.3. Nodes x_i^n and subintervals I_i^n of lengths $h_n = 2^{-n}$

The approximation to $u(x)$ is a continuous piecewise linear function $U^n(x)$ defined by the formula

$$U^n(x_j^n) = \sum_{i=1}^{j} f(x_{i-1}^n) h_n \qquad \text{for } j = 1, ..., N, \qquad (27.19)$$

where $U^n(0) = 0$. This formula gives the values of $U^n(x)$ at the nodes $x = x_j^n$ and we extend $U^n(x)$ linearly between the nodes to get the rest of the values, see Fig. 27.4

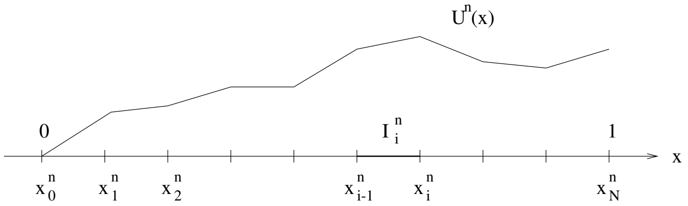


FIGURE 27.4. Piecewise linear function $U^n(x)$

We see that $U^n(x_j^n)$ is a sum of contributions $f(x_{i-1}^n)h_n$ for all intervals I_i^n with $i \leq j$. By construction,

$$U^n(x_i^n) = U^n(x_{i-1}^n) + f(x_{i-1}^n)h_n \quad \text{for } i = 1, ..., N, \tag{27.20}$$

so given the function $f(x)$, we can compute the function $U^n(x)$ by using the formula (27.20) successively with $i = 1, 2, ..., N$, where we first compute $U^n(x_1^n)$ using the value $U^n(x_0^n) = U^n(0) = 0$, then $U^n(x_2^n)$ using the value $U^n(x_1^n)$ and so on. We may alternatively use the resulting formula (27.19) involving summation, which of course just amounts to computing the sum by successively adding the terms of the sum.

The function $U^n(x)$ defined by (27.19) is thus a continuous piecewise linear function, which is computable from the nodal values $f(x_i^n)$, and we shall now motivate why $U^n(x)$ should have a good chance of being an approximation of a function $u(x)$ satisfying (27.18). If $u(x)$ is uniformly differentiable on $[0, 1]$, then

$$u(x_i^n) = u(x_{i-1}^n) + u'(x_{i-1}^n)h_n + E_u(x_i^n, x_{i-1}^n) \quad \text{for } i = 1, ..., N, \tag{27.21}$$

where $|E_u(x_i^n, x_{i-1}^n)| \leq K_u(x_i^n - x_{i-1}^n)^2 = K_u h_n^2$, and consequently

$$u(x_j^n) = \sum_{i=1}^{j} u'(x_{i-1}^n)h_n + E_h \quad \text{for } j = 1, ..., N, \tag{27.22}$$

where $|E_h| \leq K_u h_n$, since $\sum_{i=1}^{j} h_n = jh_n \leq 1$. Assuming that $u'(x) = f(x)$ for $0 < x \leq 1$, the connection between (27.20) and (27.21) and (27.19) and (27.22) becomes clear considering that the terms $E_u(x_i^n, x_{i-1}^n)$ and E_h are small. We thus expect $U^n(x_j^n)$ to be an approximation of $u(x_j^n)$ at the nodes x_j^n, and therefore $U^n(x)$ should be an increasingly accurate approximation of $u(x)$ as n increases and $h_n = 2^{-n}$ decreases.

To make this approximation idea precise, we first study the convergence of the functions $U^n(x)$ as n tends to infinity. To do this, we fix $\bar{x} \in [0, 1]$

and consider the sequence of numbers $\{U^n(\bar{x})\}_{n=1}^{\infty}$. We want to prove that this is a Cauchy sequence and thus we want to estimate $|U^n(\bar{x}) - U^m(\bar{x})|$ for $m > n$.

We begin by estimating the difference $|U^n(\bar{x}) - U^{n+1}(\bar{x})|$ for two consecutive indices n and $m = n + 1$. Recall that we used this approach in the proof of the Contraction Mapping theorem. We have

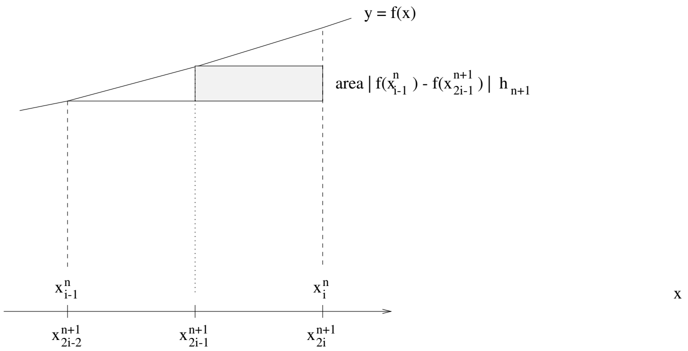


FIGURE 27.5. The difference between $U^{n+1}(x)$ and $U^n(x)$

$$U^n(x_i^n) = U^n(x_{i-1}^n) + f(x_{i-1}^n)h_n,$$

and since $x_{2i}^{n+1} = x_i^n$ and $x_{2i-2}^{n+1} = x_{i-1}^n$,

$$\begin{aligned} U^{n+1}(x_i^n) &= U^{n+1}(x_{2i}^{n+1}) = U^{n+1}(x_{2i-1}^{n+1}) + f(x_{2i-1}^{n+1})h_{n+1} \\ &= U^{n+1}(x_{i-1}^n) + f(x_{2i-2}^{n+1})h_{n+1} + f(x_{2i-1}^{n+1})h_{n+1}. \end{aligned}$$

Subtracting and setting $e_i^n = U^n(x_i^n) - U^{n+1}(x_i^n)$, we have

$$e_i^n = e_{i-1}^n + (f(x_{i-1}^n)h_n - f(x_{2i-2}^{n+1})h_{n+1} - f(x_{2i-1}^{n+1})h_{n+1}),$$

that is, since $h_{n+1} = \frac{1}{2}h_n$,

$$e_i^n - e_{i-1}^n = (f(x_{i-1}^n) - f(x_{2i-1}^{n+1}))h_{n+1}. \tag{27.23}$$

Assuming that $\bar{x} = x_j^n$ and using (27.23) and the facts that $e_0^n = 0$ and $|f(x_{i-1}^n) - f(x_{2i-1}^{n+1})| \leq L_f h_{n+1}$, we get

$$\begin{aligned}
|U^n(\bar{x}) - U^{n+1}(\bar{x})| = |e_j^n| = |\sum_{i=1}^{j}(e_i^n - e_{i-1}^n)| \\
\leq \sum_{i=1}^{j} |e_i^n - e_{i-1}^n| = \sum_{i=1}^{j} |f(x_{i-1}^n) - f(x_{2i-1}^{n+1})| h_{n+1} \\
\leq \sum_{i=1}^{j} L_f h_{n+1}^2 = \frac{1}{4} L_f h_n \sum_{i=1}^{j} h_n = \frac{1}{4} L_f \bar{x} h_n,
\end{aligned} \tag{27.24}$$

where we also used the fact that $\sum_{i=1}^{j} h_n = \bar{x}$. Iterating this estimate and using the formula for a geometric sum, we get

$$\begin{aligned}
|U^n(\bar{x}) - U^m(\bar{x})| \leq \frac{1}{4} L_f \bar{x} \sum_{k=n}^{m-1} h_k = \frac{1}{4} L_f \bar{x} (2^{-n} + ... + 2^{-m+1}) \\
= \frac{1}{4} L_f \bar{x} 2^{-n} \frac{1 - 2^{-m+n}}{1 - 2^{-1}} \leq \frac{1}{4} L_f \bar{x} 2^{-n} 2 = \frac{1}{2} L_f \bar{x} h_n,
\end{aligned}$$

that is

$$|U^n(\bar{x}) - U^m(\bar{x})| \leq \frac{1}{2} L_f \bar{x} h_n. \tag{27.25}$$

This estimate shows that $\{U^n(\bar{x})\}_{n=1}^{\infty}$ is a Cauchy sequence and thus converges to a real number. We decide, following Leibniz, to denote this real number by

$$\int_0^{\bar{x}} f(x)\, dx,$$

which thus is the limit of

$$U^n(\bar{x}) = \sum_{i=1}^{j} f(x_{i-1}^n) h_n$$

as n tends to infinity, where $\bar{x} = x_j^n$. In other words,

$$\int_0^{\bar{x}} f(x)\, dx = \lim_{n \to \infty} \sum_{i=1}^{j} f(x_{i-1}^n) h_n.$$

Letting m tend to infinity in (27.25), we can express this relation in quantitative form as follows:

$$\Big| \int_0^{\bar{x}} f(x)\, dx - \sum_{i=1}^{j} f(x_{i-1}^n) h_n \Big| \leq \frac{1}{2} L_f \bar{x} h_n.$$

At this point, we have defined the integral $\int_0^{\bar{x}} f(x)\,dx$ for a given Lipschitz continuous function $f(x)$ on $[0,1]$ as the limit of the sequence $\{U^n(\bar{x})\}_{n=1}^{\infty}$ as n tends to infinity. We can thus define a function $u:[0,1]\to\mathbb{R}$ by the formula

$$u(\bar{x}) = \int_0^{\bar{x}} f(x)\,dx \quad \text{for } \bar{x}\in[0,1]. \tag{27.26}$$

We now proceed to check that the function $u(x)$ defined in this way indeed satisfies the differential equation $u'(x)=f(x)$. We proceed in two steps. First we show that the function $u(x)$ is Lipschitz continuous on $[0,1]$ and then we show that $u'(x)=f(x)$.

Before plunging into these proofs, we need to address a subtle point. Looking back at the construction of $u(x)$, we see that we have defined $u(\bar{x})$ for $\bar{x}$ of the form $\bar{x}=x_j^n$, where $j=0,1,...,2^n$, $n=1,2,...,$. These are the rational numbers with finite decimal expansion in the base of 2, and they are *dense* in the sense that for any real number $x\in[0,1]$ and any $\epsilon>0$, there is a point of the form x_j^n so that $|x-x_j^n|\le\epsilon$. Recalling the Chapter *Real numbers*, we understand that if we can show that $u(x)$ is Lipschitz continuous on the dense set of numbers of the form x_j^n, then we can extend $u(x)$ as a Lipschitz function to the set of real numbers $[0,1]$.

We thus assume that $\bar{x}=x_j^n$ and $\bar{y}=x_k^n$ with $j>k$, and we note that

$$U^n(\bar{x})-U^n(\bar{y}) = \sum_{i=1}^{j} f(x_{i-1}^n)h_n - \sum_{i=1}^{k} f(x_{i-1}^n)h_n = \sum_{i=k+1}^{j} f(x_{i-1}^n)h_n$$

and using the triangle inequality

$$|U^n(\bar{x})-U^n(\bar{y})| \le \sum_{i=k+1}^{j} |f(x_{i-1}^n)|h_n \le M_f \sum_{i=k+1}^{j} h_n = M_f|\bar{x}-\bar{y}|,$$

where M_f is a positive constant such that $|f(x)|\le M_f$ for all $x\in[0,1]$. Letting n tend to infinity, we see that

$$u(\bar{x})-u(\bar{y}) = \int_0^{\bar{x}} f(x)\,dx - \int_0^{\bar{y}} f(x)\,dx = \int_{\bar{y}}^{\bar{x}} f(x)\,dx, \tag{27.27}$$

where of course,

$$\int_{\bar{y}}^{\bar{x}} f(x)\,dx = \lim_{n\to\infty} \sum_{i=k+1}^{j} f(x_{i-1}^n)h_n,$$

and also

$$|u(\bar{x})-u(\bar{y})| \le |\int_{\bar{y}}^{\bar{x}} f(x)\,dx| \le \int_{\bar{y}}^{\bar{x}} |f(x)|\,dx \le M_f|\bar{x}-\bar{y}|, \tag{27.28}$$

where the second inequality is the so-called *triangle inequality for integrals* to be proved in the next section. We thus have

$$|u(\bar{x}) - u(\bar{y})| \leq M_f|\bar{x} - \bar{y}|, \tag{27.29}$$

which proves the Lipschitz continuity of $u(x)$.

We now prove that the function $u(x)$ defined for $x \in [0,1]$ by the formula

$$u(x) = \int_a^x f(y)\, dy,$$

where $f : [0,1] \to \mathbb{R}$ is Lipschitz continuous, satisfies the differential equation

$$u'(x) = f(x) \quad \text{for } x \in [0,1],$$

that is

$$\frac{d}{dx}\int_0^x f(y)\, dy = f(x). \tag{27.30}$$

To this end, we choose $x, \bar{x} \in [0,1]$ with $x \geq \bar{x}$ and use (27.27) and (27.28) to see that

$$u(x) - u(\bar{x}) = \int_0^x f(z)dz - \int_0^{\bar{x}} f(y)dy = \int_{\bar{x}}^x f(y)dy,$$

and

$$\begin{aligned}
|u(x) - u(\bar{x}) - f(\bar{x})(x - \bar{x})| &= \Big|\int_{\bar{x}}^x f(y)\, dy - f(\bar{x})(x - \bar{x})\Big| \\
= \Big|\int_{\bar{x}}^x (f(y) - f(\bar{x}))\, dy\Big| &\leq \int_{\bar{x}}^x |f(y) - f(\bar{x})|\, dy \\
&\leq \int_{\bar{x}}^x L_f|y - \bar{x}|\, dy = \frac{1}{2}L_f(x - \bar{x})^2,
\end{aligned}$$

where we again used the triangle inequality for integrals. This proves that u is uniformly differentiable on $[0,1]$, and that $K_u \leq \frac{1}{2}L_f$.

Finally to prove uniqueness, we recall from (27.15) and (27.16) that a function $u : [0,1] \to \mathbb{R}$ with Lipschitz continuous derivative $u'(x)$ and $u(0) = 0$, can be represented as

$$u(\bar{x}) = \sum_{i=1}^m u'(x_{i-1})(x_i - x_{i-1}) + E_h,$$

where

$$|E_h| \leq K_u(\bar{x} - a)h.$$

Letting n tend to infinity, we find that

$$u(\bar{x}) = \int_0^{\bar{x}} u'(x)\, dx \quad \text{for } \bar{x} \in [0,1], \tag{27.31}$$

which expresses the fact that a uniformly differentiable function with Lipschitz continuous derivative is the integral of its derivative. Suppose now that $u(x)$ and $v(x)$ are two uniformly differentiable functions on $[0,1]$ satisfying $u'(x) = f(x)$, and $v'(x) = f(x)$ for $0 < x \leq 1$, and $u(0) = u_0$, $v(0) = u_0$, where $f : [0,1] \to \mathbb{R}$ is Lipschitz continuous. Then the difference $w(x) = u(x) - v(x)$ is a uniformly differentiable function on $[0,1]$ satisfying $w'(x) = 0$ for $a < x \leq b$ and $w(0) = 0$. But we just showed that

$$w(x) = \int_a^x w'(y)\, dy,$$

and thus $w(x) = 0$ for $x \in [0,1]$. This proves that $u(x) = v(x)$ for $x \in [0,1]$ and the uniqueness follows.

Recall that we proved the Fundamental Theorem for special circumstances, namely on the interval $[0,1]$ with initial value 0. We can directly generalize the construction above by replacing $[0,1]$ by an arbitrary bounded interval $[a,b]$, replacing h_n by $h_n = 2^{-n}(b-a)$, and assuming instead of $u(0) = 0$ that $u(a) = u_a$, where u_a is a given real number. We have now proved the formidable Fundamental Theorem of Calculus.

Theorem 27.1 (Fundamental Theorem of Calculus) *If $f : [a,b] \to \mathbb{R}$ is Lipschitz continuous, then there is a unique uniformly differentiable function $u : [a,b] \to \mathbb{R}$, which solves the initial value problem*

$$\begin{cases} u'(x) = f(x) & \text{for } x \in (a,b], \\ u(a) = u_a, \end{cases} \tag{27.32}$$

where $u_a \in \mathbb{R}$ is given. The function $u : [a,b] \to \mathbb{R}$ can be expressed as

$$u(\bar{x}) = u_a + \int_a^{\bar{x}} f(x)\, dx \quad \text{for } \bar{x} \in [a,b],$$

where

$$\int_0^{\bar{x}} f(x)\, dx = \lim_{n\to\infty} \sum_{i=1}^{j} f(x_{i-1}^n) h_n,$$

with $\bar{x} = x_j^n$, $x_i^n = a + ih_n$, $h_n = 2^{-n}(b-a)$. More precisely, if the Lipschitz constant of $f : [a,b] \to \mathbb{R}$ is L_f, then for $n = 1, 2, \ldots$,

$$\Big| \int_a^{\bar{x}} f(x)\, dx - \sum_{i=1}^{j} f(x_{i-1}^n) h_n \Big| \leq \frac{1}{2}(\bar{x} - a) L_f h_n. \tag{27.33}$$

Furthermore if $|f(x)| \leq M_f$ for $x \in [a,b]$, then $u(x)$ is Lipschitz continuous with Lipschitz constant M_f and $K_u \leq \frac{1}{2} L_f$, where K_u is the constant of uniform differentiability of $u : [a,b] \to \mathbb{R}$.

27.9 Comments on the notation

We can change the names of the variables and rewrite (27.26) as

$$u(x) = \int_0^x f(y)\,dy. \tag{27.34}$$

We will often use the Fundamental Theorem in the form

$$\int_a^b u'(x)\,dx = u(b) - u(a), \tag{27.35}$$

which states that the integral $\int_a^b f(x)\,dx$ is equal to the difference $u(b) - u(a)$, where $u(x)$ is a primitive function of $f(x)$. We will sometimes use the notation $[u(x)]_{x=a}^{x=b} = u(b) - u(a)$ or shorter $[u(x)]_a^b = u(b) - u(a)$, and write

$$\int_a^b u'(x)\,dx = \big[u(x)\big]_{x=a}^{x=b} = \big[u(x)\big]_a^b.$$

Sometimes the notation

$$\int f(x)\,dx,$$

without limits of integration, is used to denote a primitive function of $f(x)$. With this notation we would have for example

$$\int dx = x + C, \quad \int x\,dx = \frac{x^2}{2} + C, \quad \int x^2\,dx = \frac{x^3}{3} + C,$$

where C is a constant. We will not use this notation in this book. Note that the formula $x = \int dx$ may be viewed to express that "the whole is equal to the sum of the parts".

27.10 Alternative computational methods

Note that we might as well compute $U^n(x_i^n)$ from knowledge of $U^n(x_{i-1}^n)$, using the formula

$$U^n(x_i^n) = U^n(x_{i-1}^n) + f(x_i^n)h_n, \tag{27.36}$$

obtained by replacing $f(x_{i-1}^n)$ by $f(x_i^n)$, or

$$U^n(x_i^n) = U^n(x_{i-1}^n) + \frac{1}{2}(f(x_{i-1}^n) + f(x_i^n))h_n \tag{27.37}$$

using the mean value $\frac{1}{2}(f(x_{i-1}^n) + f(x_i^n))$. These alternatives may bring certain advantages, and we will return to them in Chapter *Numerical quadrature.* The proof of the Fundamental Theorem is basically the same with these variants and by uniqueness all the alternative constructions give the same result.

27.11 The bicyclist's speedometer

An example of a physical situation modeled by the initial value problem (27.2) is a cyclist biking along a straight line with $u(x)$ representing the position at time x, $u'(x)$ being the speed at time x and specifying the position $u(a) = u_a$ at the initial time $x = a$. Solving the differential equation (27.2) amounts to determining the position $u(x)$ of the cyclist at time $a < x \leq b$, after specifying the position at the initial time $x = a$ and knowing the speed $f(x)$ at each time x. A standard bicycle speedometer may be viewed to solve this problem, viewing the speedometer as a device which measures the instantaneous speed $f(x)$, and then outputs the total traveled distance $u(x)$. Or is this a good example? Isn't rather so that the speedometer measures the traveled distance and then reports the momentary (average) speed? To answer this question would seem to require a more precise study of how a speedometer actually works, and we urge the reader to investigate this problem.

27.12 Geometrical interpretation of the integral

In this section, we interpret the proof of the Fundamental Theorem as saying that the integral of a function is the area underneath the graph of the function. More precisely, the solution $u(\bar{x})$ given by (27.3) is equal to the area under the graph of the function $f(x)$ on the interval $[a, \bar{x}]$, see Fig. 27.6. For the purpose of this discussion, it is natural to assume that $f(x) \geq 0$.

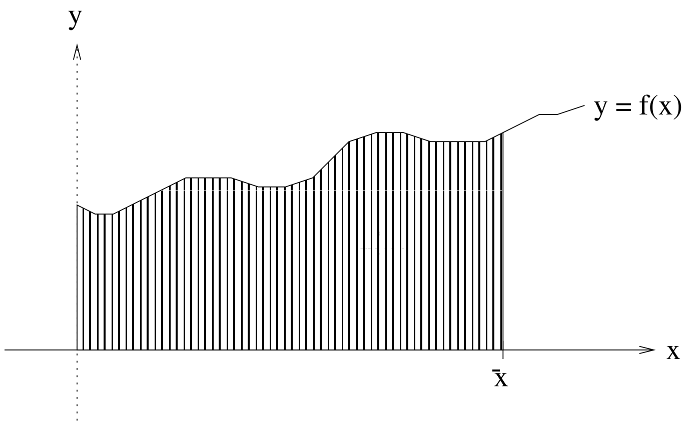


FIGURE 27.6. Area under $y = f(x)$.

Of course, we also have to explain what we mean by the area under the graph of the function $f(x)$ on the interval $[a, \bar{x}]$. To do this, we first interpret the approximation $U^n(\bar{x})$ of $u(\bar{x})$ as an area. We recall from the previous section that

$$U^n(x_j^n) = \sum_{i=1}^{j} f(x_{i-1}^n) h_n,$$

where $x_j^n = \bar{x}$. Now, we can view $f(x_{i-1}^n) h_n$ as the area of a rectangle with base h_n and height $f(x_{i-1}^n)$, see Fig. 27.7.

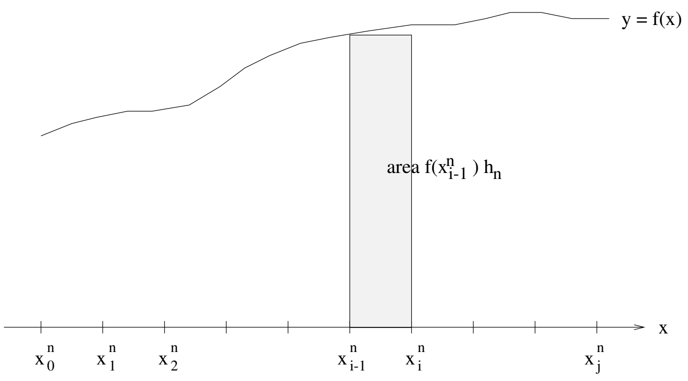


FIGURE 27.7. Area $f(x_{i-1}^n)\, h_n$ of rectangle.

We can thus interpret the sum

$$\sum_{i=1}^{j} f(x_{i-1}^n) h_n$$

as the area of a collection of rectangles which form a staircase approximation of $f(x)$, as displayed in Fig. 27.8. The sum is also referred to as a *Riemann sum*.

Intuitively, the area under the staircase approximation of $f(x)$ on $[a, \bar{x}]$, which is $U^n(\bar{x})$, will approach the area under the graph of $f(x)$ on $[a, \bar{x}]$ as n tends to infinity and $h_n = 2^{-n}(b-a)$ tends to zero. Since $\lim_{n\to\infty} U^n(\bar{x}) = u(\bar{x})$, this leads us to *define* the area under $f(x)$ on the interval $[0, \bar{x}]$ as the limit $u(\bar{x})$.

Note the logic used here: The value $U^n(\bar{x})$ represents the area under a staircase approximation of $f(x)$ on $[a, \bar{x}]$. We know that $U^n(\bar{x})$ tends to $u(\bar{x})$ as n tends to infinity, and on intuitive grounds we feel that the limit of the area under the staircase should be equal to the area under the graph of $f(x)$ on $[a, \bar{x}]$. We then simply define the area under $f(x)$ on $[a, \bar{x}]$ to be $u(\bar{x})$. By definition we thus interpret the integral of $f(x)$ on $[0, \bar{x}]$ as

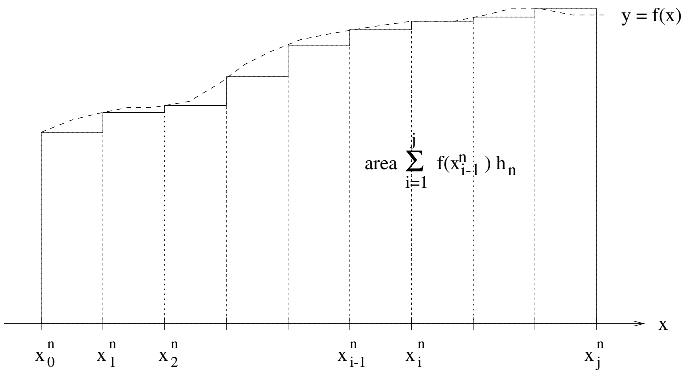


FIGURE 27.8. Area $\sum_{i=1}^{j} f(x_{i-1}^n)\, h_n$ under a staircase approximation of $f(x)$.

the area under the graph of the function $f(x)$ on $[a, \bar{x}]$. Note that *this is an interpretation*. It is not a good idea to say the integral *is* an area. This is because the integral can represent many things, such as a distance, a quantity of money, a weight, or some thing else. If we interpret the integral as an area, then we also interpret a distance, a quantity of money, a weight, or some thing else, as an area. We understand that we cannot take this interpretation to be literally true, because a distance cannot *be equal* to an area, but it can be *interpreted* as an area. We hope the reader gets the (subtle) difference.

As an example, we compute the area A under the graph of the function $f(x) = x^2$ between $x = 0$ and $x = 1$ as follows

$$A = \int_0^1 x^2\, dx = [\frac{x^3}{3}]_{x=0}^{x=1} = \frac{1}{3}.$$

This is an example of the magic of Calculus, behind its enormous success. We were able to compute an area, which in principle is the sum of very many very small pieces, without actually having to do the tedious and laborious computation of the sum. We just found a primitive function $u(x)$ of x^2 and computed $A = u(3) - u(0)$ without any effort at all. Of course we understand the telescoping sum behind this illusion, but if you don't see this, you must be impressed, right? To get a perspective and close a circle, we recall the material in Leibniz' teen-age dream in Chapter *A very short course in Calculus*.

27.13 The integral as a limit of Riemann sums

The Fundamental Theorem of Calculus states that the integral of $f(x)$ over the interval $[a,b]$ is equal to a limit of Riemann sums:

$$\int_a^b f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{2^n} f(x_{i-1}^n)h_n,$$

where $x_i^n = a + ih_n$, $h_n = 2^{-n}(b-a)$, or more precisely, for $n = 1, 2, ...,$

$$\Big| \int_a^b f(x)\,dx - \sum_{i=1}^{2^n} f(x_{i-1}^n)h_n \Big| \leq \frac{1}{2}(b-a)L_f h_n, \tag{27.38}$$

where L_f is the Lipschitz constant of f. We can thus define the integral $\int_a^b f(x)\,dx$ as a limit of Riemann sums without invoking the underlying differential equation $u'(x) = f(x)$. This approach is useful in defining integrals of functions of several variables (so-called multiple integrals like double integrals and triple integrals), because in these generalizations there is no underlying differential equation.

In our formulation of the Fundamental Theorem of Calculus, we emphasized the coupling of the integral $\int_a^x f(y)\,dy$ to the related differential equation $u'(x) = f(x)$, but as we just said, we could put this coupling in the back-ground, and define the integral as a limit of Riemann sums without invoking the underlying differential equation. This connects with the idea that the integral of a function can be interpreted as the area under the graph of the function, and will find a natural extension to multiple integrals in Chapters *Double integrals* and *Multiple integrals*.

Defining the integral as a limit of Riemann sums poses a question of uniqueness: since there are different ways of constructing Riemann sums one may ask if all limits will be the same. We will return to this question in Chapter *Numerical quadrature* and (of course) give an affirmative answer.

27.14 An Analog Integrator

James Thompson, brother of Lord Kelvin, constructed in 1876 an analog mechanical integrator based on a rotating disc coupled to a cylinder through another orthogonal disc adjustable along the radius of the first disc, see Fig. 27.10. The idea was to get around the difficulties of realizing the Analytical Engine, the mechanical digital computer envisioned by Babbage in the 1830s. Lord Kelvin tried to use a system of such analog integrators to compute different problems of practical interest including that of tide prediction, but met serious problems to reach sufficient accuracy. Similar ideas ideas were taken up by Vannevar Bush at MIT Massachussetts Institute of Technology in the 1930s, who constructed a *Differential*

Analyzer consisting of a collection of analog integrators, which was programmable to solve differential equations, and was used during the Second World War for computing trajectories of projectiles. A decade later the digital computer took over the scene, and the battle between arithmetic and geometry initiated between the Pythagorean and Euclidean schools more than 2000 years ago, had finally come an end.

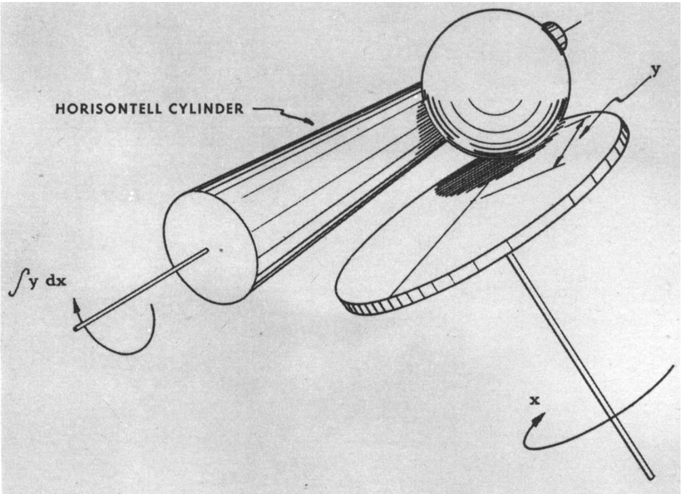


FIGURE 27.9. The principle of an Analog Integrator

Chapter 27 Problems

27.1. Determine primitive functions on $\mathbb{R}$ to (a) $(1+x^2)^{-2}2x$, (b) $(1+x)^{-99}$, (c) $(1+(1+x^3)^2)^{-2}2(1+x^3)3x^2$.

27.2. Compute the area under the graph of the function $(1+x)^{-2}$ between $x=1$ and $x=2$.

27.3. A car travels along the x-axis with speed $v(t)=t^{\frac{3}{2}}$ starting at $x=0$ for $t=0$. Compute the position of the car for $t=10$.

27.4. Carry out the proof of the Fundamental Theorem for the variations (27.36) and (27.37).

27.5. Construct a *mechanical integrator* solving the differential equation $u'(x)=f(x)$ for $x>0$, $u(0)=0$ through an analog mechanical devise. Hint: Get hold of a rotating cone and a string.

27.6. Explain the principle behind Thomson's analog integrator.

27.7. Construct a *mechanical* speedometer reporting the speed and traveled distance. Hint: Check the construction of the speedometer of your bike.

27.8. Find the solutions of the initial value problem $u'(x) = f(x)$ for $x > 0$, $u(0) = 1$, in the following cases: (a) $f(x) = 0$, (b) $f(x) = 1$, (c) $f(x) = x^r$, $r > 0$.

27.9. Find the solution to the second order initial value problem $u''(x) = f(x)$ for $x > 0$, $u(0) = u'(0) = 1$, in the following cases: (a) $f(x) = 0$, (b) $f(x) = 1$, (c) $f(x) = x^r$, $r > 0$. Explain why two initial conditions are specified.

27.10. Solve initial value problem $u'(x) = f(x)$ for $x \in (0, 2]$, $u(0) = 1$, where $f(x) = 1$ for $x \in [0, 1)$ and $f(x) = 2$ for $x \in [1, 2]$. Draw a graph of the solution and calculate $u(3/2)$. Show that $f(x)$ is not Lipschitz continuous on $[0, 2]$ and determine if $u(x)$ is Lipschitz continuous on $[0, 2]$.

27.11. The time it takes for a light beam to travel through an object is $t = \frac{d}{c/n}$, where c is the speed of light in vacuum, n is the refractive index of the object and d is its thickness. How long does it take for a light beam to travel the shortest way through the center of a glass of water, if the refractive index of the water varies as a certain function $n_w(r)$ with the distance r from the center of glass, the radius of the glass is R and the thickness and that the walls have constant thickness h and constant refractive index equal to n_g.

27.12. Assume that f and g are Lipschitz continuous on $[0, 1]$. Show that $\int_0^1 |f(x) - g(x)|dx = 0$ if and only if $f = g$ on $[0, 1]$. Does this also hold if $\int_0^1 |f(x) - g(x)|dx$ is replaced by $\int_0^1 (f(x) - g(x))dx$?

FIGURE 27.10. David Hilbert (1862-1943) at the age of 24: "A mathematical theory is not to be considered complete until you have made it so clear that you can explain it to the first man whom you meet on the street".

This is page 463
Printer: Opaque this

28
Properties of the integral

> For more than two thousand years some familiarity with mathematics has been regarded as an indipensible part of the intellectual equipment of every cultured person. Today the traditional place of mathematics in education is in great danger. Unfortunately, professional representatives of mathematics share the responsibility. The teaching of mathematics has sometimes degenerated into empty drill in problem solving, which may develop formal ability but does not lead to real understanding or to greater intellectual independence.....Teachers, students and the general public demand constructive reform, not resignation along the lines of least resistance. (Richard Courant, in Preface to What is Mathematics?, 1941)

28.1 Introduction

In this chapter, we gather together various useful properties of the integral. We may prove these properties in two ways: (i) by using the connection between the integral and the derivative and using properties of the derivative, or (ii) using that the integral is the limit of Riemann sum approximations, that is, using the area interpretation of the integral. We indicate both types of proofs to help the reader getting familiar with different aspects of the integral, and leave some of the work to the problem section.

Throughout the chapter we assume that $f(x)$ and $g(x)$ are Lipschitz continuous on the interval $[a, b]$, and we assume that

$$\sum_{i=1}^{N} f(x_{i-1}^n)h_n \qquad \text{and} \qquad \sum_{i=1}^{N} g(x_{i-1}^n)h_n$$

are Riemann sum approximations of $\int_a^b f(x)\,dx$ and $\int_a^b g(x)\,dx$ with step length $h_n = 2^{-n}(b-a)$ and $x_i^n = a + ih_n$, $i = 0, 1, ..., N = 2^n$, as in the previous chapter.

28.2 Reversing the order of upper and lower limits

So far we have defined the integral $\int_a^b f(x)\,dx$ assuming that $a \le b$, that is that the upper limit of integration b is larger than (or equal to) the lower limit a. It is useful to *extend* the definition to cases with $a > b$ by defining

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx. \tag{28.1}$$

In other words, we decide that switching the limits of integration should change the sign of an integral. As a motivation we may consider the case $f(x) = 1$ and $a > b$, and recall that $\int_b^a 1\,dx = a - b > 0$. Using the same formula with a and b interchanged, we would have $\int_a^b 1\,dx = b - a = -(a - b) = -\int_b^a 1\,dx$, which motivates the sign change under the switch of upper and lower limits. The motivation carries over to the general case using the Riemann sum approximation. Notice that here we do not *prove* anything, we simply introduce a *definition.* Of course we seek a definition which is natural, easy to remember and which allows efficient symbolic computation. The definition we chose fulfills these conditions.

EXAMPLE 28.1. We have

$$\int_2^1 2x\,dx = -\int_1^2 2x\,dx = -\big[x^2\big]_1^2 = -(4-1) = -3.$$

28.3 The whole is equal to the sum of the parts

We shall now prove that if $a \le c \le b$, then

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx. \tag{28.2}$$

One way to do this is to use the area interpretation of the integral and simply notice that the area under $f(x)$ from a to b should be equal to the sum of the area under $f(x)$ from a to c and the area under $f(x)$ from c to b.

We can also give an alternative proof using that that $\int_a^b f(x)\,dx = u(b)$, where $u(x)$ satisfies $u'(x) = f(x)$ for $a \le x \le b$, and $u(a) = 0$. Letting now $w(x)$ satisfy $w'(x) = f(x)$ for $c \le x \le b$, and $w(c) = u(c)$, we have by uniqueness that $w(x) = u(x)$ for $c \le x \le b$, and thus

$$u(b) = w(b) = u(c) + \int_c^b f(y)\,dy = \int_a^c f(y)\,dy + \int_c^b f(y)\,dy,$$

which is the desired result.

Example 28.2. We have

$$\int_0^2 x\,dx = \int_0^1 x\,dx + \int_1^2 x\,dx,$$

which expresses the identity

$$2 = \left(\frac{1}{2}\right) + \left(2 - \frac{1}{2}\right).$$

Note that by the definition (28.1), (28.2) actually holds for any a, b and c.

28.4 Integrating piecewise Lipschitz continuous functions

A functions is said to be *piecewise Lipschitz continuous* on a finite interval $[a, b]$ if $[a, b]$ can be divided up into a finite number of sub-intervals on which the function is Lipschitz continuous, allowing the function to have jumps at the ends of the subintervals, see Fig. 28.1. We now extend (in the obvious way) the definition of the integral $\int_a^b f(x)\,dx$ to a piecewise Lipschitz continuous function $f(x)$ on an interval $[a, b]$ starting with the case of two subintervals with thus $f(x)$ Lipschitz continuous separately on two adjoining intervals $[a, c]$ and $[c, b]$, where $a \le c \le b$. We define

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx,$$

which obviously fits with (28.2). The extension is analogous for several subintervals with the integral over the whole interval being the sum of the integrals over the subintervals.

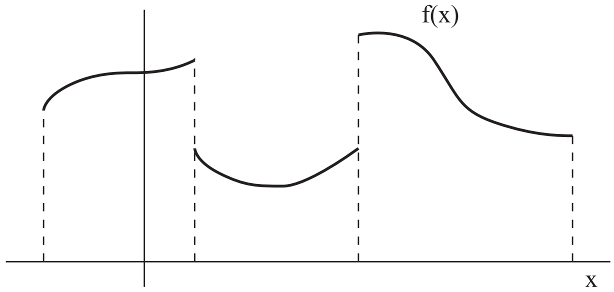


FIGURE 28.1. A piecewise Lipschitz continuous function.

28.5 Linearity

We shall now prove the following property of *linearity* of the integral: If α and β are real numbers then,

$$\int_a^b (\alpha f(x) + \beta g(x))\, dx = \alpha \int_a^b f(x)\, dx + \beta \int_a^b g(x)\, dx. \tag{28.3}$$

With $\alpha = \beta = 1$ this property expresses that the area (from a to b) underneath the sum of two functions is equal to the sum of the areas underneath each function. Further, with $g(x) = 0$ and $\alpha = 2$ say, the property expresses that the area under the function $2f(x)$ is equal to 2 times the area under the function $f(x)$.

More generally, the linearity of the integral is inherited from the linearity of the Riemann sum approximation, which we may express as

$$\sum_{i=1}^{N} (\alpha f(x_{i-1}^n) + \beta g(x_{i-1}^n)) h_n = \alpha \sum_{i=1}^{N} f(x_{i-1}^n) h_n + \beta \sum_{i=1}^{N} g(x_{i-1}^n) h_n, \tag{28.4}$$

and which follows from basic rules for computing with real numbers.

A differential equation proof goes as follows: Define

$$u(x) = \int_a^x f(y)\, dy \quad \text{and} \quad v(x) = \int_a^x g(y)\, dy, \tag{28.5}$$

that is, $u(x)$ is a primitive function of $f(x)$ satisfying $u'(x) = f(x)$ for $a < x \leq b$ and $u(a) = 0$, and $v(x)$ is a primitive function of $g(x)$ satisfying $v'(x) = g(x)$ for $a < x \leq b$ and $v(a) = 0$. Now, the function $w(x) = \alpha u(x) + \beta v(x)$ is a primitive function of the function $\alpha f(x) + \beta g(x)$, since by the linearity of the derivative, $w'(x) = \alpha u'(x) + \beta v'(x) = \alpha f(x) + \beta g(x)$, and $w(a) = \alpha u(a) + \beta v(a) = 0$. Thus, the left hand side of (28.3) is equal to $w(b)$, and since $w(b) = \alpha u(b) + \beta v(b)$, the desired equality follows from setting $x = b$ in (28.5).

EXAMPLE 28.3. We have

$$\int_0^b (2x + 3x^2)\, dx = 2\int_0^b x\, dx + 3\int_0^b x^2\, dx = 2\frac{b^2}{2} + 3\frac{b^3}{3} = b^2 + b^3.$$

28.6 Monotonicity

The *monotonicity* property of the integral states that if $f(x) \geq g(x)$ for $a \leq x \leq b$, then

$$\int_a^b f(x)\, dx \geq \int_a^b g(x)\, dx. \tag{28.6}$$

This is the same as stating that if $f(x) \geq 0$ for $x \in [a, b]$, then

$$\int_a^b f(x)\, dx \geq 0, \tag{28.7}$$

which follows from the fact that all Riemann sum approximations $\sum_{i=1}^{j} f(x_{i-1}^n)h_n$ of $\int_a^b f(x)\, dx$ are all non-negative if $f(x) \geq 0$ for $x \in [a, b]$.

28.7 The triangle inequality for integrals

We shall now prove the following *triangle inequality for integrals*:

$$\Big|\int_a^b f(x)\, dx\Big| \leq \int_a^b |f(x)|\, dx, \tag{28.8}$$

stating that moving the absolute value inside the integral increases the value (or leaves the value unchanged). This property follows from applying the usual triangle inequality to Riemann sum approximations to get

$$\Big|\sum_{i=1}^{N} f(x_{i-1}^n)h_n\Big| \leq \sum_{i=1}^{N} \Big|f(x_{i-1}^n)\Big| h_n$$

and then passing to the limit. Evidently there may be cancellations on the left hand side if $f(x)$ has changes sign, while on the right hand side we always add nonnegative contributions, making the right hand side at least as big as the left hand side.

Another proof uses the monotonicity as follows: Apply (28.7) to the function $|f| - f \geq 0$ to obtain

$$\int_a^{\bar{x}} f(x)\, dx \leq \int_a^{\bar{x}} |f(x)|\, dx.$$

Replacing f by the function $-f$ we obtain

$$-\int_a^{\bar{x}} f(x)\,dx = \int_a^{\bar{x}} (-f(x))\,dx \le \int_a^{\bar{x}} |-f(x)|\,dx = \int_a^{\bar{x}} |f(x)|\,dx,$$

which proves the desired result.

28.8 Differentiation and integration are inverse operations

The Fundamental Theorem expresses that integration and differentiation are *inverse operations* in the sense that first integrating and then differentiating, or first differentiating and then integrating, gives the net result of doing nothing! We make this clear by repeating a part of the proof of the Fundamental Theorem to prove that if $f : [a,b] \to \mathbb{R}$ is Lipschitz continuous then for $x \in [a,b]$,

$$\frac{d}{dx}\int_a^x f(y)\,dy = f(x). \tag{28.9}$$

In other words, integrating a function $f(x)$ and then differentiating the primitive function, gives back the function $f(x)$. Surprise? We illustrate in Fig. 28.2. To properly understand the equality (28.9), it is important to

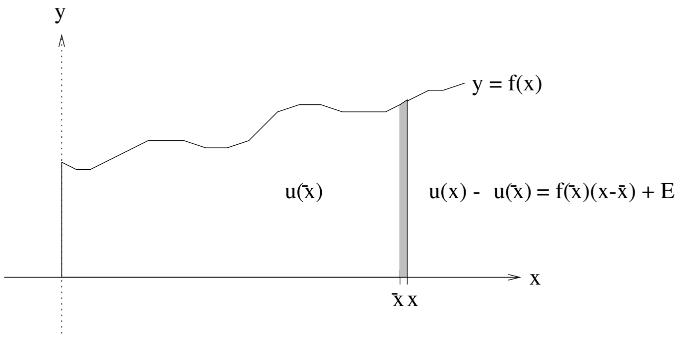


FIGURE 28.2. The derivative of $\int_0^x f(y)\,dy$ at $x = \bar{x}$ is $f(\bar{x})$: $|E| \le \frac{1}{2}L_f|x-\bar{x}|^2$.

realize that $\int_a^x f(y)\,dy$ is a function of x and thus depends on x. The area under the function f from a to x, of course depends on the upper limit x.

We may express (28.9) in words as follows: Differentiating an integral with respect to the upper limit of integration gives the value of the integrand at the upper limit of integration.

To prove (28.9) we now choose x and $\bar{x}$ in $[a,b]$ with $x \geq \bar{x}$, and use (28.2) to see that

$$u(x) - u(\bar{x}) = \int_a^x f(z)\,dz - \int_a^{\bar{x}} f(y)dy = \int_{\bar{x}}^x f(y)\,dy$$

so that

$$\begin{aligned}
|u(x) - u(\bar{x}) - f(\bar{x})(x-\bar{x})| &= \Big|\int_{\bar{x}}^x f(y)\,dy - f(\bar{x})(x-\bar{x})\Big| \\
&= \Big|\int_{\bar{x}}^x (f(y) - f(\bar{x}))\,dy\Big| \\
&\leq \int_{\bar{x}}^x |f(y) - f(\bar{x})|\,dy \\
&\leq \int_{\bar{x}}^x L_f|y-\bar{x}|\,dy = \frac{1}{2}L_f(x-\bar{x})^2.
\end{aligned}$$

This proves that $u(x)$ is uniformly differentiable on $[a,b]$ with derivative $u'(x) = f(x)$ and constant $K_u \leq \frac{1}{2}L_f$.

We also note that (28.1) implies

$$\frac{d}{dx}\int_x^a f(y)\,dy = -f(x). \tag{28.10}$$

which we may express in words as: Differentiating an integral with respect to the lower limit of integration gives minus the value of the integrand at the lower limit of integration.

Example 28.4. We have

$$\frac{d}{dx}\int_0^x \frac{1}{1+y^2}\,dy = \frac{1}{1+x^2}.$$

Example 28.5. Note that we can combine (28.9) with the Chain Rule:

$$\frac{d}{dx}\int_0^{x^3} \frac{1}{1+y^2}\,dy = \frac{1}{1+(x^3)^2}\frac{d}{dx}\left(x^3\right) = \frac{3x^2}{1+x^6}.$$

28.9 Change of variables or substitution

We recall that the Chain rule tells us how to differentiate the composition of two functions. The analogous property of the integral is known as the *change of variables*, or *substitution* formula and plays an important role. For example, it can be used to compute many integrals analytically. The idea is that an integral may be easier to compute analyticlly if we change scales in the independent variable.

Let now $g : [a, b] \to I$, be uniformly differentiable on an interval $[a, b]$, where I is an interval, and let $f : I \to \mathbb{R}$ be Lipschitz continuous. Typically, g is strictly increasing (or decreasing) and maps $[a, b]$ onto I, so that $g : [a, b] \to I$ corresponds to a change of scale, but more general situations are allowed. The *change of variables* formula reads

$$\int_a^x f(g(y))g'(y)\,dy = \int_{g(a)}^{g(x)} f(z)\,dz \quad \text{for } x \in [a, b]. \tag{28.11}$$

This is also called *substitution* since the left hand side $L(x)$ is formally obtained by in the right hand hand side $H(x)$ setting $z = g(y)$ and replacing dz by $g'(y)\,dy$ motivated by the relation

$$\frac{dz}{dy} = g'(y),$$

and noting that as y runs from a to x then z runs from $g(a)$ to $g(x)$.

To verify (28.11), we now prove that $H'(x) = L'(x)$ and use the fact that $H(a) = L(a) = 0$ and the uniqueness of the integral. The Chain rule and (28.9) imply that

$$H'(x) = f(g(x))\,g'(x).$$

Further,

$$L'(x) = f(g(x))\,g'(x),$$

which thus proves the desired equality.

We now give a two examples. We will meet many more examples below.

Example 28.6. To integrate

$$\int_0^2 (1 + y^2)^{-2} 2y\,dy$$

we make the observation that

$$\frac{d}{dy}(1 + y^2) = 2y.$$

Thus, if we set $z = g(y) = 1 + y^2$, then applying (28.11) noting that $g(0) = 1$ and $g(2) = 5$ and formally $dz = 2ydy$, we have that

$$\int_0^2 (1 + y^2)^{-2} 2y\,dy = \int_0^2 (g(y))^{-2} g'(y)\,dy = \int_1^5 z^{-2}\,dz.$$

Now, the right hand integral can easily be evaluated:

$$\int_1^5 z^{-2}\,dz = [-z^{-1}]_{z=1}^{z=5} = -\big(\frac{1}{5} - 1\big),$$

and thus

$$\int_0^2 (1 + y^2)^{-2} 2y\,dy = \frac{4}{5}.$$

EXAMPLE 28.7. We have setting $y = g(x) = 1 + x^4$ noting that then formally $dy = g'(x)dx = 4x^3dx$ and $g(0) = 1$ and $g(1) = 2$, to get

$$\int_0^1 (1+x^4)^{-1/2}x^3\,dx = \frac{1}{4}\int_0^1 (g(x))^{-1/2}g'(x)\,dx = \frac{1}{4}\int_1^2 y^{-1/2}\,dy$$
$$=\frac{1}{2}[y^{1/2}]_1^2 = \frac{\sqrt{2}-1}{2}.$$

28.10 Integration by parts

We recall that the Product rule is a basic property of the derivative, showing how to compute the derivative of a product of two functions. The corresponding formula for integration is called *integration by parts*. The formula is

$$\int_a^b u'(x)v(x)\,dx = u(b)v(b) - u(a)v(a) - \int_a^b u(x)v'(x)\,dx. \tag{28.12}$$

The formula follows by applying the Fundamental Theorem to the function $w(x) = u(x)v(x)$, in the form

$$\int_a^b w'(x)\,dx = u(b)v(b) - u(a)v(a),$$

together with the product formula $w'(x) = u'(x)v(x)+u(x)v'(x)$ and (28.3). Below we often write

$$u(b)v(b) - u(a)v(a) = \Big[u(x)v(x)\Big]_{x=a}^{x=b},$$

and we can thus state the formula for integration by parts as

$$\int_a^b u'(x)v(x)\,dx = \Big[u(x)v(x)\Big]_{x=a}^{x=b} - \int_a^b u(x)v'(x)\,dx. \tag{28.13}$$

This formula is very useful and we will use many times below.

EXAMPLE 28.8. Computing

$$\int_0^1 4x^3(1+x^2)^{-3}\,dx$$

by guessing at a primitive function for the integrand would be a pretty daunting task. However we can use integration by parts to compute the integral. The trick here is to realize that

$$\frac{d}{dx}(1+x^2)^{-2} = -4x(1+x^2)^{-3}.$$

If we rewrite the integral as

$$\int_0^1 x^2 \times 4x(1+x^2)^{-3}\,dx$$

then we can apply integration by parts with $u(x) = x^2$ and $v'(x) = 4x(1+x^2)^{-3}$, so $u'(x) = 2x$ and $v(x) = -(1+x^2)^{-2}$, to get

$$\begin{aligned}\int_0^1 4x^3(1+x^2)^{-3}\,dx &= \int_0^1 u(x)v'(x)\,dx \\ &= [x^2(-(1+x^2)^{-2})]_{x=0}^{x=1} - \int_0^1 2x(-(1+x^2)^{-2})\,dx \\ &= -\frac{1}{4} - \int_0^1 (-(1+x^2)^{-2})2x\;dx.\end{aligned}$$

To do the remaining integral, we use the substitution $z = 1 + x^2$ with $dz = 2x\,dx$ to get

$$\begin{aligned}\int_0^1 4x^3(1+x^2)^{-3}\,dx &= -\frac{1}{4} + \int_1^2 z^{-2}\,dz \\ &= -\frac{1}{4} + \big[-z^{-1}\big]_{z=1}^{z=2} = -\frac{1}{4} - \frac{1}{2} + 1 = \frac{1}{4}.\end{aligned}$$

28.11 The Mean Value Theorem

The *Mean Value theorem* states that if $u(x)$ is a differentiable function on $[a, b]$ then there is a point $\bar{x}$ in (a, b) such that the slope $u'(\bar{x})$ of the tangent of the graph of $u(x)$ at $\bar{x}$ is equal to the slope of the secant line, or chord, connecting the points $(a, u(a))$ and $(b, u(b))$. In other words,

$$\frac{u(b) - u(a)}{b - a} = u'(\bar{x}). \tag{28.14}$$

This is geometrically intuitive, see Fig. 28.3, and expresses that the average velocity over $[a, b]$ is equal to momentary velocity $u'(\bar{x})$ at some intermediate point $\bar{x} \in [a, b]$. To get from the point $(a, u(a))$ to the point $(b, u(b))$, f has to "bend" around in such a way that the tangent becomes parallel to the secant line at least at one point.

Assuming that $u'(x)$ is Lipschitz continuous on $[a, b]$, we shall now prove that there is a real number $\bar{x} \in [a, b]$ such that

$$u(b) - u(a) = (b - a)u'(\bar{x})$$

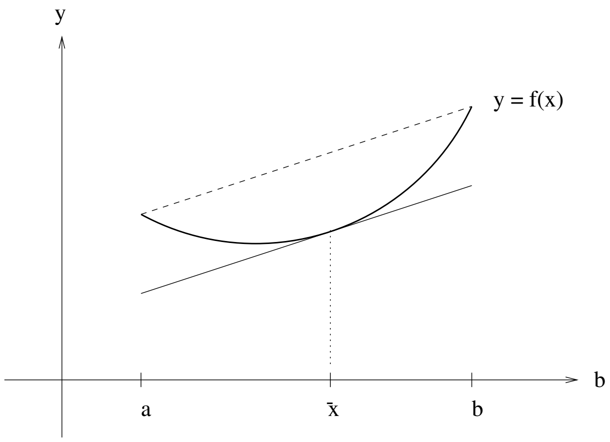


FIGURE 28.3. Illustration of the Mean Value Theorem.

which is equivalent to (28.14). The proof is based on the formula

$$u(b) = u(a) + \int_a^b u'(x)\, dx \qquad (28.15)$$

which holds if $u(x)$ is uniformly differentiable on $[a, b]$. If for all $x \in [a, b]$, we would have

$$\frac{u(b) - u(a)}{b - a} > u'(x),$$

then we would have (explain why)

$$u(b) - u(a) = \int_b^a \frac{u(b) - u(a)}{b - a}\, dx > \int_b^a u'(x)\, dx = u(b) - u(a),$$

which is a contradiction. Arguing in the same way, we conclude it is also impossible that

$$\frac{u(b) - u(a)}{b - a} < u'(x)$$

for all $x \in [a, b]$. So there must be numbers c and d in $[a, b]$ such that

$$u'(c) \leq \frac{u(b) - u(a)}{b - a} \leq u'(d).$$

Since $u'(x)$ is Lipschitz continuous for $x \in [a, b]$, it follows by the Intermediate Value theorem that there is a number $\bar{x} \in [a, b]$ such that

$$u'(\bar{x}) = \frac{u(b) - u(a)}{b - a}.$$

We have now proved:

Theorem 28.1 (Mean Value theorem) *If $u(x)$ is uniformly differentiable on $[a,b]$ with Lipschitz continuous derivative $u'(x)$, then there is a (at least one) $\bar{x} \in [a,b]$, such that*

$$u(b) - u(a) = (b-a)u'(\bar{x}). \tag{28.16}$$

The Mean Value Theorem is often written in terms of an integral by setting $f(x) = u'(x)$ in (28.16), which gives

Theorem 28.2 (Mean Value theorem for integrals) *If $f(x)$ is Lipschitz continuous on $[a,b]$, then there is some $\bar{x} \in [a,b]$ such that*

$$\int_a^b f(x)\,dx = (b-a)f(\bar{x}). \tag{28.17}$$

The Mean Value theorem turns out to be very useful in several ways. To illustrate, we discuss two results that can be proved easily using the Mean Value Theorem.

28.12 A function with non-negative derivative is non-decreasing

The first result says that the sign of the derivative of a function indicates whether the function is increasing or decreasing in value as the input increases. More precisely, the Mean Value Theorem implies that if $f'(x) \geq 0$ for all $x \in [a,b]$ then $f(b) \geq f(a)$. Moreover if $x_1 \leq x_2$ are in $[a,b]$, then $f(x_1) \leq f(x_2)$. A function with this property is said to be *non-decreasing* on $[a,b]$. If in fact $f'(x) > 0$ for all $x \in (a,b)$, then $f(x_1) < f(x_2)$ for $x_1 < x_2$ in $[a,b]$ (with strict inequalities), and we say that $f(x)$ is *(strictly) increasing* on the interval $[a,b]$. Corresponding statements hold if $f'(x) \leq 0$ and $f'(x) < 0$, with non-decreasing and (strictly) increasing replaced with *non-increasing* and *(strictly) decreasing*, respectively.

28.13 A function with zero derivative is constant

As a particular consequence of the preceeding section, we conclude that if $f'(x) = 0$ for all $x \in [a,b]$, so that $f(x)$ is both non-increasing and non-decreasing on $[a,b]$, then $f(x)$ is constant on $[a,b]$. Thus, a function with derivative vanishing everywhere is a constant function.

28.14 Bounded derivative implies Lipschitz continuity (again)

As a second consequence of the Mean Value Theorem, we give an alternate, and shorter, proof that a function with a Lipschitz continuous derivative is Lipschitz continuous. Assume that $u : [a,b] \to \mathbb{R}$ has a Lipschitz continuous derivative $u'(x)$ on $[a,b]$ satisfying $|u'(x)| \leq M$ for $x \in [a,b]$. By the Mean Value theorem, we have

$$|u(x) - u(\bar{x})| \leq M|x - \bar{x}| \quad \text{for } x, \bar{x} \in [a,b].$$

We conclude that $u(x)$ is Lipschitz continuous on $[a,b]$ with Lipschitz constant $M = \max_{x\in[a,b]} |u'(x)|$.

28.15 Taylor's theorem

In earlier chapters, we analyzed a linear approximation to a function u,

$$u(x) \approx u(\bar{x}) + u'(\bar{x})(x - \bar{x}), \tag{28.18}$$

as well as a quadratic approximation

$$u(x) \approx u(\bar{x}) + u'(\bar{x})(x - \bar{x}) + \frac{u''(\bar{x})}{2}(x - \bar{x})^2. \tag{28.19}$$

These approximations are very useful tools for dealing with nonlinear functions. *Taylor's theorem*, invented by Brook Taylor (1685-1731), see Fig. 28.5, generalizes these approximations to any degree.Taylor sided up with Newton in a long scientific fight with the Leibnizians about "who's best in Calculus?".

Theorem 28.3 (Taylor's theorem) *If $u(x)$ is $n+1$ times differentiable on the interval I with $u^{(n+1)}$ Lipschitz continuous, then for x, $\bar{x} \in I$, we have*

$$u(x) = u(\bar{x}) + u'(\bar{x})(x - \bar{x}) + \cdots + \frac{u^{(n)}(\bar{x})}{n!}(x - \bar{x})^n + \int_{\bar{x}}^{x} \frac{(x-y)^n}{n!} u^{(n+1)}(y)\, dy. \tag{28.20}$$

The polynomial

$$P_n(x) = u(\bar{x}) + u'(\bar{x})(x - \bar{x}) + \cdots + \frac{u^{(n)}(\bar{x})}{n!}(x - \bar{x})^n$$

is called the *Taylor polynomial*, or *Taylor expansion*, of $u(x)$ at $\bar{x}$ of degree n,

FIGURE 28.4. Brook Taylor, inventor of the Taylor expansion: "I am the best".

The term
$$R_n(x) = \int_{\bar{x}}^{x} \frac{(x-y)^n}{n!}\, u^{(n+1)}(y)\, dy$$
is called the *remainder term* of order n. We have for $x \in I$,
$$u(x) = P_n(x) + R_n(x).$$
It follows directly that
$$\left(\frac{d^k}{dx^k}\right) P_n(\bar{x}) = \left(\frac{d^k}{dx^k}\right) u(\bar{x}) \quad \text{for } k = 0, 1 \cdots, n.$$
Thus Taylor's theorem gives a polynomial approximation $P_n(x)$ of degree n of a given function $u(x)$, such that the derivatives of order $\leq n$ of $P_n(x)$ and $u(x)$ agree at $x = \bar{x}$.

EXAMPLE 28.9. The Taylor polynomial of order 2 at $x = 0$ for $u(x) = \sqrt{1+x}$ is given by
$$P_2(x) = 1 + \frac{1}{2} - \frac{1}{8}x^2,$$
since $u(0) = 1$, $u'(0) = \frac{1}{2}$, and $u''(0) = -\frac{1}{4}$.

The proof of Taylor's theorem is a wonderful application of integration by parts, discovered by Taylor. We start by noting that Taylor's theorem with $n = 0$ is just the Fundamental Theorem

$$u(x) = u(\bar{x}) + \int_{\bar{x}}^{x} u'(y)\, dy,$$

Using that $\frac{d}{dy}(y - x) = 1$, we get integrating by parts

$$\begin{aligned}
u(x) &= u(\bar{x}) + \int_{\bar{x}}^{x} u'(y)\, dy \\
&= u(\bar{x}) + \int_{\bar{x}}^{x} \frac{d}{dy}(y - x) u'(y)\, dy \\
&= u(\bar{x}) + [(y - x)u'(y)]_{y=\bar{x}}^{y=x} - \int_{\bar{x}}^{x} (y - x)u''(y)\, dy \\
&= u(\bar{x}) + (x - \bar{x})u'(\bar{x}) + \int_{\bar{x}}^{x} (x - y)u''(y)\, dy,
\end{aligned}$$

which is Taylor's theorem with $n = 1$. Continuing in this manner, integrating by parts, using the notation $k_n(y) = (y - x)^n/n!$, and noting that for $n \geq 1$

$$\frac{d}{dy} k_n(y) = k_{n-1}(y),$$

we get

$$\begin{aligned}
&\int_{\bar{x}}^{x} \frac{(x - y)^{n-1}}{(n-1)!} u^{(n)}(y)\, dy = (-1)^{n-1} \int_{\bar{x}}^{x} k_{n-1}(y) u^{(n)}(y)\, dy \\
&= (-1)^{n-1} \int_{\bar{x}}^{x} \frac{d}{dy} k_n(y) u^{(n)}(y)\, dy \\
&= [(-1)^{n-1} k_n(y) u^{(n)}(y)]_{y=\bar{x}}^{y=x} - (-1)^{n-1} \int_{\bar{x}}^{x} k_n(y) u^{(n+1)}(y)\, dy \\
&= \frac{u^{(n)}(\bar{x})}{n!}(x - \bar{x})^n + \int_{\bar{x}}^{x} \frac{(x - y)^n}{n!} u^{(n+1)}(y)\, dy.
\end{aligned}$$

This proves Taylor's theorem.

EXAMPLE 28.10. We compute a fourth order polynomial approximation to $f(x) = \frac{1}{1-x}$ near $x = 0$. We have

$$\begin{aligned}
f(x) &= \frac{1}{1 - x} &&\Longrightarrow f(0) = 1, \\
f'(x) &= \frac{1}{(1 - x)^2} &&\Longrightarrow f'(0) = 1, \\
f''(x) &= \frac{2}{(1 - x)^3} &&\Longrightarrow f''(0) = 2, \\
f'''(x) &= \frac{6}{(1 - x)^4} &&\Longrightarrow f'''(0) = 6, \\
f''''(x) &= \frac{24}{(1 - x)^5} &&\Longrightarrow f''''(0) = 24,
\end{aligned}$$

and therefore

$$P_4(x) = 1 + 1(x-0)^1 + \frac{2}{2}(x-0)^2 + \frac{6}{6}(x-0)^3 + \frac{24}{24}(x-0)^4$$
$$= 1 + x + x^2 + x^3 + x^4.$$

We plot the function and the polynomial in Fig. 28.5. Characteristically,

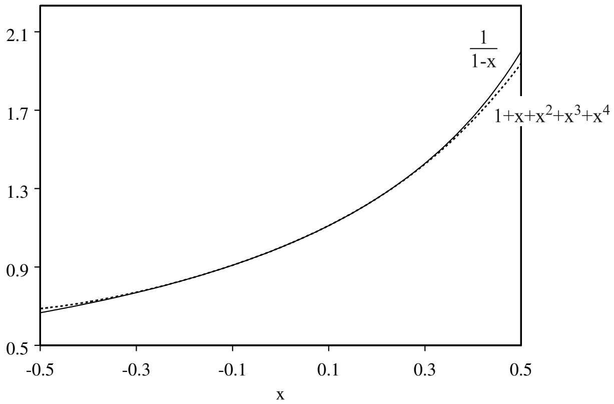


FIGURE 28.5. Plots of $f(x) = 1/(1-x)$ and its Taylor polynomial $1 + x + x^2 + x^3 + x^4$.

the Taylor polynomial is a very accurate approximation near the $\bar{x}$ but the error becomes larger as x moves further away from $\bar{x}$.

28.16 October 29, 1675

On October 29, 1675, Leibniz got a bright idea while sitting at his desk in Paris. He wrote "Utile erit scribit $\int$ pro omnia", which translates to "It is useful to write $\int$ instead of omnia". This is the moment when the modern notation of calculus was created. Earlier than this date, Leibniz had been working with a notation based on a, l and "omnia" which represented in modern notation dx, dy and $\int$ respectively. This notation resulted in formulas like

$$\text{omn.}l = y, \quad \text{omn.}yl = \frac{y^2}{2}, \quad \text{omn.}xl = x\text{omn.}l - \text{omn.omn.}la,$$

where "omn.", short for omnia, indicated a discrete sum and l and a denoted increments of finite size (often $a = 1$). In the new notation, these formulas became

$$\int dy = y, \quad \int y\,dy = \frac{y^2}{2}, \quad \int x\,dy = xy - \int y\,dx. \tag{28.21}$$

This opened up the possibility of dx and dy being arbitrarily small and the sum being replaced by the "integral".

28.17 The hodometer

The Romans constructed many roads to keep the Empire together and the need of measuring distances between cities and traveled distance on the road, became very evident. For this purpose the *Hodometer* was constructed by Vitruvius XXXX, see Fig. 28.6. For each turn of the wagon wheel, the vertical gear is shifted one step, and for each turn of the vertical gear the horisontal gear is shifted one step. The horisontal gear has a set of holes with one stone in each, and for each shift one stone drops down to a box under the wagon; at the end of the day one computes the number of stones in the box, and the device is so calibrated that this number is equal to the number of traveled miles. Evidently, one may view the hodometer as a kind of simple analog integrator.

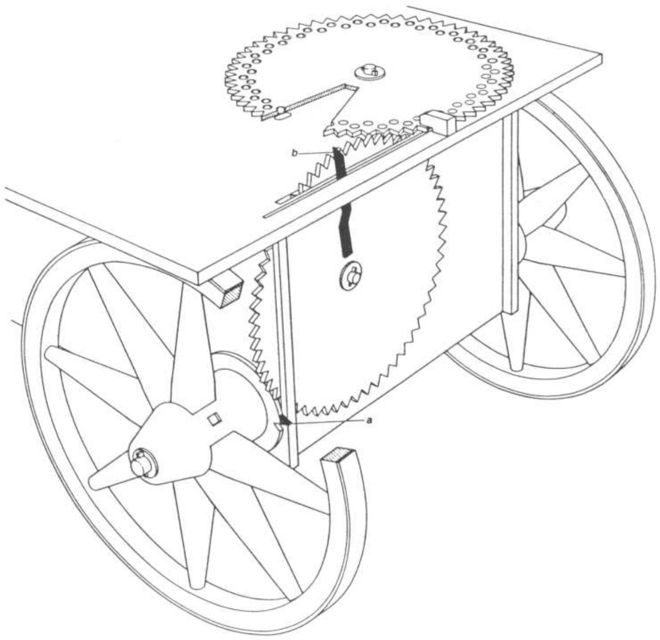


FIGURE 28.6. The principle of the Hodometer

Chapter 28 Problems

28.1. Compute the following integrals: a) $\int_0^1 (ax+bx^2)dx$, b) $\int_{-1}^1 |x|dx$, c) $\int_{-1}^1 |x-1|dx$, d) $\int_{-1}^1 |x+a|dx$, e) $\int_{-1}^1 (x-a)^{10}dx$.

28.2. Compute the following integrals by integration by parts. Verify that you get the same result by directly finding the primitive function. a) $\int_0^1 x^2dx = \int_0^1 x \cdot x dx$, b) $\int_0^1 x^3dx = \int_0^1 x \cdot x^2dx$, c) $\int_0^1 x^3dx = \int_0^1 x^{3/2} \cdot x^{3/2}dx$, d) $\int_0^1 (x^2-1)dx = \int_0^1 (x+1)\cdot(x-1)dx$.

28.3. For computing the integral $\int_0^1 x(x-1)^{1000}dx$, what would you rather do; find the primitive function directly or integrate by parts?

28.4. Compute the following integrals: a) $\int_{-1}^2 (2x-1)^7dx$, b) $\int_0^1 f'(7x)dx$, c) $\int_{-10}^{-7} f'(17x+5)dx$.

28.5. Compute the integral $\int_0^1 x(x^2-1)^{10}dx$ in two ways, first by integration by parts, then by a clever substitution using the chain rule.

28.6. Find Taylor polynomials at $\overline{x}$ of the following functions: a) $f(x) = x$, $\overline{x} = 0$, b) $f(x) = x + x^2 + x^3$, $\overline{x} = 1$, c) $f(x) = \sqrt{\sqrt{x+1}+1}$, $\overline{x} = 0$.

28.7. Find a Taylor expansion of the function $f(x) = x^r - 1$ around a suitable choice of $\overline{x}$, and use the result to compute the limit $\lim_{x\to 1} \frac{x^r-1}{x-1}$. Compare this to using l'Hopital's rule to compute the limit. Can you see the connection between the two methods?

28.8. Motivate the basic properties of linearity and subinterval additivity of the integral using the area interpretation of the integral.

28.9. Prove the basic properties of linearity and subinterval additivity of the integral using that the integral is a limit of discrete sums together with basic properties of discrete sums.

28.10. Make sense out of Leibniz formulas (28.21). Prove, as did Leibniz, the second from a geometrical argument based on computing the area of a right-angled triangle by summing thin slices of variable height y and thickness dy, and the third from computing similarly the area of a rectangle as the sum of the two parts below and above a curve joining two opposite corners of the rectangle.

28.11. Prove the following variant of Taylor's theorem: If $u(x)$ is $n+1$ times differentiable on the interval I, with $u^{(n+1)}(x)$ Lipschitz continuous, then for $\bar{x} \in I$, we have

$$\begin{aligned} u(x) = u(\bar{x}) + u'(\bar{x})(x-\bar{x}) + \cdots + \frac{u^{(n)}(\bar{x})}{n!}(x-\bar{x})^n \\ + \frac{u^{(n+1)}(\hat{x})}{(n+1)!}(x-\bar{x})^{n+1} \end{aligned}$$

where $\hat{x} \in [\bar{x}, x]$. Hint: Use the Mean Value theorem for integrals.

28.12. Prove that if $x = f(y)$ with inverse function $y = f^{-1}(x)$, and $f(0) = 0$, then

$$\int_0^{\bar{y}} f(y)\, dy = \bar{y}\bar{x} - \int_0^{\bar{x}} f^{-1}(x)\, dx.$$

Compare with (28.21) Hint: use integration by parts.

28.13. Show that $x \mapsto F(x) = \int_0^x f(x)dx$ is Lipschitz continuous on $[0, a]$ with Lipschitz constant L_F if $|f(x)| \leq L_F$ for $x \in [0, a]$.

28.14. Why can we think of the primitive function as being "nicer" than the function itself?

28.15. Under what conditions is the following generalized integration by parts formula valid

$$\int_I \frac{d^n f}{dx^n}\varphi dx = (-1)^n \int_I f \frac{d^n \varphi}{dx^n} dx, \ \ n = 0, 1, 2, \ldots?$$

28.16. Show the following inequality:

$$|\int_I u(x)v(x)\ dx| \leq \sqrt{\int_I u^2(x)\ dx}\ \sqrt{\int_I v^2(x)\ dx},$$

which is referred to as *Cauchy's inequality.* Hint: Let $\overline{u} = u/\sqrt{\int_I u^2(x)\ dx}$, $\overline{v} = v/\sqrt{\int_I v^2(x)\ dx}$, and show that $|\int_I \overline{u}(x)\overline{v}(x)\ dx| \leq 1$ by considering the expression $\int_I(\overline{u}(x) - \int_I \overline{u}(y)\overline{v}(y)\ dy\ \overline{v}(x))\ dx$. Would it be helpful to use the notation $(u, v) = \int_I u(x)v(x)\ dx$, and $\|u\| = \sqrt{\int_I u^2(x)\ dx}$?

28.17. Show that if v is Lipschitz continuous on the bounded interval I and $v = 0$ at one of the endpoints of the interval, then

$$\|v\|_{L^2(I)} \leq C_I\|v'\|_{L^2(I)},$$

for some constant C_I, where the so-called $L^2(I)$ norm of a function v is defined as $\|v\|_{L^2(I)} = \sqrt{\int_I v^2(x)dx}$. What is the value of the constant? Hint: Express v in terms of v' and use the result from the previous problem.

28.18. Check that the inequality from the previous problem holds for the following functions on $I = [0, 1]$: a) $v(x) = x(1 - x)$, b) $v(x) = x^2(1 - x)$, c) $v(x) = x(1 - x)^2$,

28.19. Prove quadratic convergence of Newton's method (25.5) for computing a root $\bar{x}$ of the equation $f(x) = 0$ using Taylor's theorem. Hint: Use the fact that $x_{i+1} - \bar{x} = x_i - \bar{x} + \frac{f(x_i) - f(\bar{x})}{f'(x_i)}$ and Taylor's theorem to see that $f(x_i) - f(\bar{x}) = f'(x_i)(x_i - \bar{x}) + \frac{1}{2}f''(\tilde{x}_i)(x_i - \bar{x})^2$ for some $\tilde{x}_i \approx x_i$.

28.20. Prove (28.3) from (28.4).

29
The logarithm $\log(x)$

Nevertheless technicalities and detours should be avoided, and the presentation of mathematics should be just as free from emphasis on routine as from forbidding dogmatism, which refuses to disclose motive or goal and which is an unfair obstacle to honest effort. (R. Courant)

29.1 The definition of $\log(x)$

We return to the question of the existence of a primitive function of $f(x) = 1/x$ for $x > 0$ posed above. Since the function $f(x) = 1/x$ is Lipschitz continuous on any given interval $[a, b]$ with $0 < a < b$, we know by the Fundamental Theorem that there is a unique function $u(x)$ which satisfies $u'(x) = 1/x$ for $a \leq x \leq b$ and takes on a specific value at some point in $[a, b]$, for example $u(1) = 0$. Since $a > 0$ may be chosen as small as we please and b as large as we please, we may consider the function $u(x)$ to be defined for $x > 0$. We now define the *natural logarithm* $\log(x)$ (or $\ln(x)$) for $x > 0$ as the primitive function $u(x)$ of $1/x$ vanishing for $x = 1$, i.e., $\log(x)$ satisfies

$$\frac{d}{dx}(\log(x)) = \frac{1}{x} \quad \text{for } x > 0, \quad \log(1) = 0. \tag{29.1}$$

Using the definition of the integral, we may express $\log(x)$ as an integral:

$$\log(x) = \int_1^x \frac{1}{y}\, dy \quad \text{for } x > 0. \tag{29.2}$$

In the next chapter we shall use this formula to compute approximations of $\log(x)$ for a given $x > 0$ by computing approximations of the corresponding integral. We plot $\log(x)$ in Fig. 29.1.

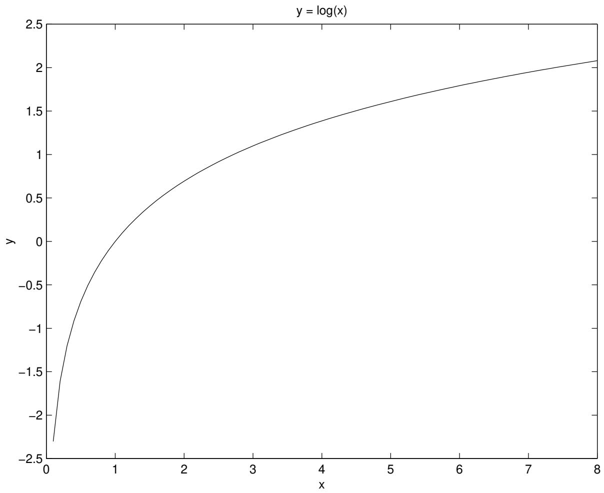


FIGURE 29.1. Plot of $\log(x)$.

29.2 The importance of the logarithm

The logarithm function $\log(x)$ is a basic function in science, simply because it solves a basic differential equation, and thus occurs in many applications. More concretely, the logarithm has some special properties that made previous generations of scientists and engineers use the logarithm itensely, including memorizing long tables of its values. The reason is that one can compute products of real numbers by adding logarithms of real numbers, and thus the operation of multiplication can be replaced by the simpler operation of addition. The *slide rule* is an analog computing device built on this principle, that used to be sign of the engineer visible in the waist-pocket, recall Fig. 1.5. Today the modern computer has replaced the slide rule and does not use logarithms to multiply real numbers. However, the first computer, the mechanical Difference Machine by Babbage from the 1830s, see Fig. 1.2, was used for computing accurate tables of values of the logarithm. The logarithm was discovered by John Napier and presented in *Mirifici logarithmorum canonis descriptio* in 1614. A illuminating citation from the foreword is given in Fig. 29.2.

FIGURE 29.2. Napier, Inventor of the Logarithm: "Seeing there is nothing (right well-beloved Students of the Mathematics) that is so troublesome to mathematical practice, nor that doth more molest and hinder calculators, than the multiplications, divisions, square and cubical extractions of great numbers, which besides the tedious expense of time are for the most part subject to many slippery errors, I began therefore to consider in my mind by what certain and ready art I might remove those hindrances. And having thought upon many things to this purpose, I found at length some excellent brief rules to be treated of (perhaps) hereafter. But amongst all, none more profitable than this which together with the hard and tedious multiplications, divisions, and extractions of roots, doth also cast away from the work itself even the very numbers themselves that are to be multiplied, divided and resolved into roots, and putteth other numbers in their place which perform as much as they can do, only by addition and subtraction, division by two or division by three.

29.3 Properties: $\log(ab) = \log(a) + \log(b)$ and $\log(a^r) = r\log(a)$

We now derive the basic properties of the logarithm function $\log(x)$ using (29.1) or (29.2). We first note that $u(x) = \log(x)$ is *strictly increasing* for $x > 0$, because $u'(x) = 1/x$ is positive for $x > 0$. This can be seen in Fig. 29.1. Recalling (**??**), we conclude that for $a, b > 0$,

$$\int_a^b \frac{dy}{y} = \log(b) - \log(a).$$

Next we note that the Chain rule implies that for any constant $a > 0$

$$\frac{d}{dx}(\log(ax) - \log(x)) = \frac{1}{ax} \cdot a - \frac{1}{x} = 0,$$

and consequently $\log(ax) - \log(x)$ is constant for $x > 0$. Since $\log(1) = 0$, we see by setting $x = 1$ that the constant value is equal to $\log(a)$, and so

$$\log(ax) - \log(x) = \log(a) \quad \text{for } x > 0.$$

Choosing $x = b > 0$, we thus obtain the following fundamental relation satisfied by the logarithm $\log(x)$:

$$\log(ab) = \log(a) + \log(b) \quad \text{for } a, b > 0 \tag{29.3}$$

We can thus compute the logarithm of the product of two numbers by adding the logarithms of the two numbers. We already indicated that this is the principle of the slide rule or using a table of logarithms for multiplying two numbers. More precisely (as first proposed by Napier), to multiply two numbers a and b we first find their logarithms $\log(a)$ and $\log(b)$ from the table, then add them to get $\log(ab)$ using the formula (29.3), and finally we find from the table which real number has the logarithm equal to $\log(ab)$, which is equal to the desired product ab. Clever, right?

The formula (29.3) implies many things. For example, choosing $b = 1/a$, we get

$$\log(a^{-1}) = -\log(a) \quad \text{for } a > 0. \tag{29.4}$$

Choosing $b = a^{n-1}$ with $n = 1, 2, 3, ..,$ we get

$$\log(a^n) = \log(a) + \log(a^{n-1}),$$

so that by repeating this argument

$$\log(a^n) = n\log(a) \quad \text{for } n = 1, 2, 3, ... \tag{29.5}$$

By (29.4) the last equality holds also for $n = -1, -2, ...$

More generally, we have for any $r \in \mathbb{R}$ and $a > 0$,

$$\log(a^r) = r\log(a). \tag{29.6}$$

We prove this using the change of variables $x = y^r$ with $dx = ry^{r-1}dy$:

$$\log(a^r) = \int_1^{a^r} \frac{1}{x}\,dx = \int_1^a \frac{ry^{r-1}}{y^r}\,dy = r\int_1^a \frac{1}{y}\,dy = r\log(a).$$

Finally, we note that $1/x$ also has a primitive function for $x < 0$ and for $a, b > 0$, setting $y = -x$,

$$\begin{aligned}\int_{-a}^{-b} \frac{dy}{y} = \int_a^b \frac{-dx}{-x} = \int_a^b \frac{dx}{x} &= \log(b) - \log(a)\\ &= \log(-(-b)) - \log(-(-a)).\end{aligned}$$

Accordingly, we may write for any $a \neq 0$ and $b \neq 0$ that have the same sign,

$$\int_a^b \frac{dx}{x} = \log(|b|) - \log(|a|). \tag{29.7}$$

It is important to understand that (29.7) does *not* hold if a and b have opposite signs.

Chapter 29 Problems

29.1. Prove that $\log(4) > 1$ and $\log(2) \geq 1/2$.

29.2. Prove that

$$\begin{aligned} \log(x) &\to \infty \quad \text{as } x \to \infty, \\ \log(x) &\to -\infty \quad \text{as } x \to 0^+. \end{aligned}$$

Hint: Using that $\log(2) \geq 1/2$ it follows from (29.5) that $\log(2^n)$ tends to infinity as n tends to infinity.

29.3. Give an alternative proof of (29.3) using that

$$\log(ab) = \int_1^{ab} \frac{1}{y}dy = \int_1^a \frac{1}{y}dy + \int_a^{ab} \frac{1}{y}dy = \log(a) + \int_a^{ab} \frac{1}{y}dy,$$

and changing the variable y in the last integral to $z = ay$.

29.4. Prove that $\log(1+x) \leq x$ for $x > 0$, and that $\log(1+x) < x$ for $x \neq 0$ and $x > -1$. Hint: Differentiate.

29.5. Show using the Mean Value theorem, that $\log(1+x) \leq x$ for $x > -1$. Can prove this directly from the definition of the logarithm by sketching the area under the graph?

29.6. Prove that $\log(a) - \log(b) = \log(\frac{a}{b})$ for $a, b > 0$.

29.7. Write down the Taylor polynomial of order n for $\log(x)$ at $x = 1$.

29.8. Find a primitive function of $\frac{1}{x^2-1}$. Hint: use that $\frac{1}{x^2-1} = \frac{1}{(x-1)(x+1)} = \frac{1}{2}(\frac{1}{x-1} - \frac{1}{x+1})$.

29.9. Prove that $\log(x^r) = r\log(x)$ for $r = \frac{p}{q}$ rational by using (29.5) cleverly.

29.10. Solve the initial value problem $u'(x) = 1/x^a$ for $x > 0$, $u(1) = 0$, for values of the exponent a close to 1. Plot the solutions. Study for which values of a the solution $u(x)$ tends to infinity when x tends to infinity.

29.11. Solve the following equations: (a) $\log(x^2) + \log(3) = \log(\sqrt{x}) + \log(5)$, (b) $\log(7x) - 2\log(x) = \log(3)$, (c) $\log(x^3) - \log(x) = \log(7) - \log(x^2)$.

29.12. Compute the derivatives of the following functions: a) $f(x) = \log(x^3+6x)$, b) $f(x) = \log(\log(x))$, c) $f(x) = \log(x + x^2)$, d) $f(x) = \log(1/x)$, e) $f(x) = x\log(x) - x$.

30
Numerical quadrature

> "And I know it *seems* easy", said Piglet to himself, "but it isn't *everyone* who could do it". (House at the Pooh Corner, Milne)
>
> Errare humanum est.

30.1 Computing integrals

In some cases, we can compute a primitive function (or antiderivative or integral) of a given function analytically, that is we can give give a formula for the primitive function in terms of known functions. For example we can give a formula for a primitive function of a polynomial as another polynomial. We will return in Chapter *Techniques of integration* to the question of finding analytical formulas for primitive functions of certain classes of functions. The Fundamental Theorem states that any given Lipschitz continuous function has a primitive function, but does not give any analytical formula for the primitive function. The logarithm,

$$\log(x) = \int_1^x \frac{dy}{y}, \quad \text{where } x > 0,$$

is the first example of this case we have encountered. We know that the logarithm function $\log(x)$ exists for $x > 0$, and we have derived some of its properties indirectly through its defining differential equation, but the question remains how to determine the value of $\log(x)$ for a given $x > 0$. Once we have solved this problem, we may add $\log(x)$ to a list of "elementary" functions that we can play with. Below we will add to this list the exponential function, the trigonometric functions, and other more exotic functions.

This situation is completely analogous to solving algebraic equations for numbers. Some equations have rational roots and in that case, we feel that we can solve the equation "exactly" by analytic (symbolic) computation. We have a good understanding of rational numbers, even when they have infinite decimal expansions, and we can determine their values, or the pattern in the decimal expansion, with a finite number of arithmetic operations. But most equations have irrational roots with infinite, nonrepeating decimal expansions that we can only approximate to a desired level of accuracy in practice. Likewise in a situation in which a function is known only as a primitive function of a given function, the best we can do is to seek to compute its values approximately to a desired level of accuracy. One way to compute values of such a function is through the definition of the integral as a Riemann sum. This is known as *numerical quadrature* or *numerical integration*, and we now explore this possibility.

Suppose thus that we want to compute the integral

$$\int_a^b f(x)\,dx, \tag{30.1}$$

where $f : [a,b] \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant L_f. If we can give a formula for the primitive function $F(x)$ of $f(x)$, then the integral is simply $F(b) - F(a)$. If we cannot give a formula for $F(x)$, then we turn to the Fundamental Theorem and compute an approximate the value of the integral using a Riemann sum approximation

$$\int_a^b f(x)\,dx \approx \sum_{i=1}^{N} f(x_{i-1}^n)h_n, \tag{30.2}$$

where $x_i^n = a + ih_n$, $h_n = 2^{-n}(b-a)$, and $N = 2^n$ describes a uniform partition of $[a,b]$, with the *quadrature error*

$$Q_n = \Big|\int_a^b f(x)\,dx - \sum_{i=1}^{N} f(x_{i-1}^n)h_n\Big| \le \frac{b-a}{2}L_f h_n, \tag{30.3}$$

which tends to zero as we increase the number of steps and $h_n \to 0$. Put another way, if we desire the value of the integral to within a *tolerance* $TOL > 0$ and we know the Lipschitz constant L_f, then we will have $Q_n \le TOL$ if the mesh size h_n satisfies the *stopping criterion*

$$h_n \le \frac{2\,TOL}{(b-a)L_f}. \tag{30.4}$$

We refer to the Riemann sum approximation (30.2), compare also with Fig. 30.1, as the *rectangle rule*, which is the simplest method for approximating an integral among many possible methods. The search for more sophisticated methods for approximating an integral is driven by consideration of the *computational cost* associated to computing the approximation. The cost is typically measured in terms of time because there is a limit on the time we are willing to wait for a solution. In the rectangle rule, the computer spends most of the time evaluating the function f and since each step requires one evaluation of f, the cost is determined ultimately by the number of steps. Considering the cost leads to the optimization problem of trying to compute an approximation of a given accuracy at a relatively low cost.

To reduce the cost, we may construct more sophisticated methods for approximating integrals. But even if we restrict ourselves to the rectangle rule, we can introduce variations that can lower the computational cost of computing an approximation. There are two quantities that we can vary: the point at which we evaluate the function on each interval and the size of the intervals. To understand how these changes could help, consider the illustration of the rectangle rule in Fig. 30.1. Here f varies quite a bit on part

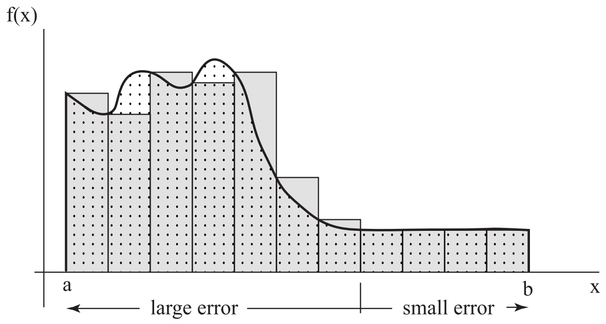


FIGURE 30.1. An illustration of the rectangle rule.

of $[a, b]$ and is fairly constant on another part. Consider the approximation to the area under f on the first subinterval on the left. By evaluating f at the left-hand point on the subinterval, we clearly overestimate the area to the maximum degree possible. Choosing to evaluate f at some point inside the subinterval would likely give a better approximation. The same is true of the second subinterval, where choosing the left-hand point clearly leads to an underestimate of the area. On the other hand, consider the approximations to the area in the last four subintervals on the right. Here f is nearly constant and the approximation is very accurate. In fact, we could approximate the area underneath f on this part of $[a, b]$ using one rectangle rather than four. In other words, we would get just as accurate an approximation by using one large subinterval instead of four subintervals. This would cost four times less.

So we generalize the rectangle rule to allow non-uniform partitions and different points at which to evaluate f. We choose a partition $a = x_0 < x_1 < x_2 \cdots < x_N = b$ of $[a, b]$ into N subintervals $I_j = [x_{j-1}, x_j]$ of lengths $h_j = x_j - x_{j-1}$ for $j = 1, ..., N$. Note that N can be any integer and the subintervals may vary in size. By the Mean Value theorem for integrals there is $\bar{x}_j \in I_j$ such that

$$\int_{x_{j-1}}^{x_j} f(x)\,dx = f(\bar{x}_j)h_j, \tag{30.5}$$

and thus we have

$$\int_a^b f(x)\,dx = \sum_{i=1}^N \int_{x_{j-1}}^{x_j} f(x)\,dx = \sum_{j=1}^N f(\bar{x}_j)h_j.$$

Since the $\bar{x}_j$ are not known in general, we replace $\bar{x}_j$ by a given point $\hat{x}_j \in I_j$. For example, in the original method we use the left end-point $\hat{x}_j = x_{j-1}$, but we could choose the right end-point $\hat{x}_j = x_j$ or the mid-point $\hat{x}_j = \frac{1}{2}(x_{j-1} + x_j)$. We then get the approximation

$$\int_a^b f(x)\,dx \approx \sum_{j=1}^N f(\hat{x}_j)h_j. \tag{30.6}$$

We call

$$\sum_{j=1}^N f(\hat{x}_j)h_j \tag{30.7}$$

a *quadrature* formula for computing the integral $\int_a^b f(x)\,dx$. We recall that we refer to (30.7) as a *Riemann sum*. The quadrature formula is characterized by the *quadrature points* $\hat{x}_j$ and the *weights* h_j. Note that if $f(x) = 1$ for all x then the quadrature formula is exact and we conclude that $\sum_{j=1}^N h_j = b - a$.

We now estimate the *quadrature error*

$$Q_h = \Big| \int_a^b f(x)\,dx - \sum_{j=1}^N f(\hat{x}_j)h_j \Big|,$$

where the subscript h refers to the sequence of step sizes h_j. Recalling (30.5), we can do this by estimating the error over each subinterval and then summing. We have

$$\Big| \int_{x_{j-1}}^{x_j} f(x)\,dx - f(\hat{x}_j)h_j \Big| = |f(\bar{x}_j)h_j - f(\hat{x}_j)h_j| = h_j|f(\bar{x}_j) - f(\hat{x}_j)|.$$

We assume that $f'(x)$ is Lipschitz continuous on $[a, b]$. The Mean Value theorem implies that for $x \in [x_{j-1}, x_j]$,

$$f(x) = f(\hat{x}_j) + f'(y)(x - \hat{x}_j),$$

for some $y \in [x_{j-1}, x_j]$. Integrating over $[x_{j-1}, x_j]$ we obtain

$$\left| \int_{x_{j-1}}^{x_j} f(x)\, dx - f(\hat{x}_j) h_j \right| \leq \max_{y \in I_j} |f'(y)| \int_{x_{j-1}}^{x_j} |x - \hat{x}_j|\, dx.$$

To simplify the sum on the right, we use the fact that

$$\int_{x_{j-1}}^{x_j} |x - \hat{x}_j|\, dx$$

is maximized if $\hat{x}_j$ is the left (or right) endpoint, in which case

$$\int_{x_{j-1}}^{x_j} (x - x_{j-1})\, dx = \frac{1}{2} h_j^2.$$

We find that

$$\left| \int_{x_{j-1}}^{x_j} f(x)\, dx - f(\hat{x}_j) h_j \right| \leq \frac{1}{2} \max_{y \in I_j} |f'(y)| h_j^2.$$

Summing, we conclude

$$Q_h = \left| \int_a^b f(x)\, dx - \sum_{j=1}^N f(\hat{x}_j) h_j \right| \leq \frac{1}{2} \sum_{j=1}^N \left(\max_{y \in I_j} |f'(y)|\, h_j \right) h_j. \tag{30.8}$$

This generalizes the estimate of the Fundamental Theorem to non-uniform partitions. We can see that (30.8) implies that Q_h tends to zero as the maximal step size tends to zero by estimating further:

$$Q_h \leq \frac{1}{2} \max_{[a,b]} |f'| \sum_{j=1}^N h_j \max_{1 \leq j \leq N} h_j = \frac{1}{2}(b - a) \max_{[a,b]} |f'| \max_{1 \leq j \leq N} h_j. \tag{30.9}$$

So Q_h tends to zero at the same rate that $\max h_j$ tends to zero.

30.2 The integral as a limit of Riemann sums: again

We now return to the (subtle) question posed at the end of the Chapter The Integral: Will all limits of Rieman sum approximations (as the maximal subinterval tends to zero) of a certain integral be the same? We recall that we defined the integral using a particular uniform partition and we now ask if any limit of non-uniform partions will be the same. The affirmative answer follows from the last statement of the previous section: The quadrature error Q_h tends to zero as $\max h_j$ tends to zero, under the assumption that $\max_{[a,b]} |f'|$ is finite, that is $|f'(x)|$ is bounded on $[a, b]$. This proves the uniqueness of limits of Riemann sum approximations of a certain integral as the maximal subinterval tends to zero, under the assumption that the derivative of the integrand is bounded. This assumption can naturally be relaxed to assuming that the integrand is Lipschitz continuous. We sum up:

Theorem 30.1 *The limit (as the maximal subinterval tends to zero) of Riemann sum approximations of an integral of a Lipschitz continuous function, is unique.*

30.3 The midpoint rule

We now analyze the quadrature formula in which the quadrature point is chosen to be the midpoint of each subinterval, $\hat{x}_j = \frac{1}{2}(x_{j-1} + x_j)$. It turns out that this choice gives a formula that is more accurate than any other rectangle rule on a given mesh provided f has a Lipschitz continuous second derivative. Taylor's theorem implies that for $x \in [x_{j-1}, x_j]$,

$$f(x) = f(\hat{x}_j) + f'(\hat{x}_j)(x - \hat{x}_j) + \frac{1}{2} f''(y)(x - \hat{x}_j)^2,$$

for some $y \in [x_{j-1}, x_j]$ if we assume that f'' is Lipschitz continuous. We argue as above by integrating over $[x_{j-1}, x_j]$. Now however we use the fact that

$$\int_{x_{j-1}}^{x_j} (x - \hat{x}_j)\, dx = \int_{x_{j-1}}^{x_j} (x - (x_j + x_{j-1})/2)\, dx = 0$$

which holds only when $\hat{x}_j$ is the midpoint of $[x_{j-1}, x_j]$. This gives

$$\begin{aligned} \Big| \int_{x_{j-1}}^{x_j} f(x)\, dx - f(\hat{x}_j) h_j \Big| &\le \frac{1}{2} \max_{y \in I_j} |f''(y)| \int_{x_{j-1}}^{x_j} (x - \hat{x}_j)^2\, dx \\ &\le \frac{1}{24} \max_{y \in I_j} |f''(y)| h_j^3. \end{aligned}$$

Now summing the errors on each subinterval, we obtain the following estimate on the total error

$$Q_h \le \frac{1}{24} \sum_{j=1}^{N} \left(\max_{y \in I_j} |f''(y)|\, h_j^2 \right) h_j. \qquad (30.10)$$

To understand the claim that this formula is more accurate than any other rectangle rule, we estimate further

$$Q_h \leq \frac{1}{24}(b-a) \max_{[a,b]} |f''| \max h_j^2$$

which says that the error decreases as $\max h_j$ decreases like $\max h_j^2$. Compare this to the general result (30.9), which says that the error decreases like $\max h_j$ for general rectangle rules. If we halve the step size $\max h_j$ then in a general rectangle rule the error decreases by a factor of two but in the midpoint rule the error decreases by a factor of *four*. We say that the midpoint rule converges at a *quadratic* rate while the general rectangle rule converges at a *linear* rate.

We illustrate the accuracy of these methods and the error bounds by approximating

$$\log(4) = \int_1^4 \frac{dx}{x} \approx \sum_{j=1}^{N} \frac{h_j}{\hat{x}_j}$$

both with the original rectangle rule with $\hat{x}_j$ equal to the left-hand endpoint x_{j-1} of each subinterval and the midpoint rule. In both cases, we use a constant stepsize $h_i = (4-1)/N$ for $i = 1, 2, \cdots, N$. It is straightforward to evaluate (30.8) and (30.10) because $|f'(x)| = 1/x^2$ and $|f''(x)| = 2/x^3$ are both decreasing functions. We show the results for four different values of N.

	Rectangle rule		Midpoint rule	
N	True error	Error bound	True error	Error bound
25	.046	.049	.00056	.00056
50	.023	.023	.00014	.00014
100	.011	.011	.000035	.000035
200	.0056	.0057	.0000088	.0000088

These results show that the error bounds (30.8) and (30.10) can give quite accurate estimates of the true error. Also note that the midpoint rule is much more accurate than the general rectangle rule on a given mesh and moreover the error in the midpoint rule goes to zero quadratically with the error decreasing by a factor of 4 each time the number of steps is doubled.

30.4 Adaptive quadrature

In this section, we consider the optimization problem of trying to compute an approximation of an integral to within a given accuracy at a relatively low cost. To simplify the discussion, we use the original rectangle rule with $\hat{x}_j$ equal to the left-hand endpoint x_{j-1} of each subinterval to compute the approximation. The optimization problem becomes to compute an approximation with error smaller than a given tolerance TOL using the least number of steps. Since we do not know the error of the approximation, we use the quadrature estimate (30.8) to estimate the error. The optimization problem is therefore to find a partition $\{x_j\}_{j=0}^N$ using the smallest number of points N that satisfies the stopping criterion

$$\sum_{j=1}^{N} \left(\max_{y \in I_j} |f'(y)| \, h_j \right) h_j \leq \text{TOL}. \tag{30.11}$$

This equation suggests that we should adjust or *adapt* the stepsizes h_j depending on the size of $\max_{I_j} \, |f'|$. If $\max_{I_j} \, |f'|$ is large, then h_j should be small, and vice versa. Trying to find such an optimized partition is refered to as *adaptive* quadrature, because we seek a partition suitably adapted to the nature of the integrand $f(x)$.

There are several possible strategies for finding such a partition and we consider two here.

In the first strategy, or adaptive algorithm, we estimate the sum in (30.11) as follows

$$\sum_{j=1}^{N} \left(\max_{I_j} |f'| \, h_j \right) h_j \leq (b-a) \max_{1 \leq j \leq N} \left(\max_{I_j} |f'| \, h_j \right),$$

where we use the fact that $\sum_{j=1}^N h_j = b - a$. It follows that (**??**) is satisfied if the steps are chosen by

$$h_j = \frac{\text{TOL}}{(b-a) \max\limits_{I_j} |f'|} \quad \text{for } j = 1, ..., N. \tag{30.12}$$

In general, this corresponds to a nonlinear equation for h_j since $\max_{I_j} |f'|$ depends on h_j.

We apply this adaptive algorithm to the computation of $\log(b)$ and obtain the following results

TOL	b	Steps	Approximate Area	Error
.05	4.077	24	1.36	.046
.005	3.98	226	1.376	.0049
.0005	3.998	2251	1.38528	.0005
.00005	3.9998	22501	1.3861928	.00005

The reason b varies slightly in these results is due to the strategy we use to implement (30.12). Namely, we specify the tolerance and then search for the value of N that gives the closest b to 4.

We plot the sequence of mesh sizes for TOL = .01 in Fig. 30.2, where the adaptivity is plainly visible. In contrast, if we compute with a uniform mesh, we find using (30.11) that we need $N = 9/\text{TOL}$ points to guarantee an accuracy of TOL. For example, this means using 900 points to guarantee an accuracy of .01, which is significantly more than needed for the adapted mesh.

The second adaptive algorithm is based on an *equidistribution of error* in which the steps h_j are chosen so that the contribution to the error from each sub-interval is roughly equal. Intuitively, this should lead to the least number of intervals since the largest error reduction is gained if we subdivide the interval with largest contribution to the error. In this case, we estimate the sum on the left-hand side of (30.11) by

$$\sum_{j=1}^{N} \left(\max_{I_j} |f'|\, h_j \right) h_j \leq N \max_{1\leq j\leq N} \left(\max_{I_j} |f'|\, h_j^2 \right)$$

and determine the steps h_j by

$$h_j^2 = \frac{\text{TOL}}{N \max_{I_j} |f'|} \quad \text{for } j = 1, ..., N. \tag{30.13}$$

As above, we have to solve a nonlinear equation for h_j, now with the additional complication of the explicit presence of the total number of steps N.

We implement (30.13) to compute of $\log(b)$ with $b \approx 4$ and obtain the following results:

TOL	b	Steps	Approximate Area	Error
.05	4.061	21	1.36	.046
.005	4.0063	194	1.383	.005
.0005	3.9997	1923	1.3857	.0005
.00005	4.00007	19220	1.38626	.00005

We plot the sequence of step sizes for TOL $= .01$ in (30.2). We see that at every tolerance level, the second adaptive strategy (30.13) gives the same accuracy at $x_N \approx 4$ as (30.12) while using fewer steps. It thus seems that the second algorithm is more efficient.

We compare the efficiency of the two adaptive algorithms by estimating the number of steps N required to compute $\log(x)$ to a given tolerance TOL in each case. We begin by noting that the equality

$$N = \frac{h_1}{h_1} + \frac{h_2}{h_2} + \cdots + \frac{h_N}{h_N},$$

implies that, assuming $x_N > 1$,

$$N = \int_1^{x_N} \frac{dy}{h(y)},$$

where $h(y)$ is the piecewise constant *mesh function* with the value $h(s) = h_j$ for $x_{j-1} < s \leq x_j$. In the case of the second algorithm, we substitute the

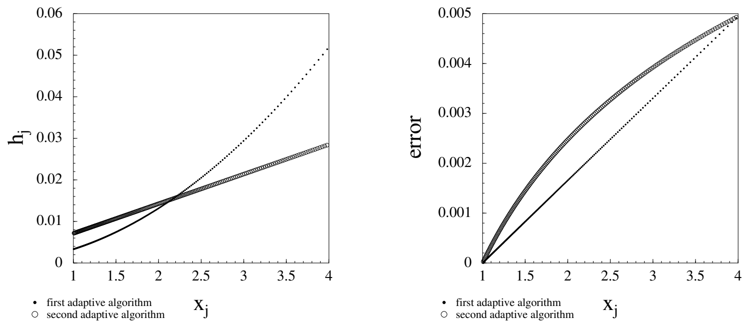


FIGURE 30.2. On the left, we plot the step sizes generated by two adaptive algorithms for the integration of log(4) using TOL = .01. On the right, we plot the errors of the same computations versus x.

value of h given by (30.13) into the integral to get, recalling that $f(y) = 1/y$ so that $f'(y) = -1/y^2$,

$$N \approx \frac{\sqrt{N}}{\sqrt{\mathrm{TOL}}} \int_1^{x_N} \frac{dy}{y},$$

or

$$N \approx \frac{1}{\mathrm{TOL}} \left(\log(x_N)\right)^2. \tag{30.14}$$

Making a similar analysis of the first adaptive algorithm, we get

$$N \approx \frac{x_{N-1}}{\mathrm{TOL}} \big(1 - \frac{1}{x_N}\big). \tag{30.15}$$

We see that in both cases, N is inversely proportional to TOL. However, the number of steps needed to reach the desired accuracy using the first adaptive algorithm increases much more quickly as x_N increases than the number needed by the second algorithm, i.e. at a linear rate as opposed to a logarithmic rate. Note that the case $0 < x_N < 1$ may be reduced to the case $x_N > 1$ by replacing x_N by $1/x_N$ since $\log(x) = -\log(1/x)$.

If we use (30.12) or (30.13) to choose the steps h_j over the interval $[a, x_N]$, then of course the quadrature error over any smaller interval $[a, x_i]$ with $i \le N$, is also smaller than TOL. For the first algorithm (30.12), we can actually show the stronger estimate

$$\left| \int_a^{x_i} f(y)\, dy - \sum_{j=1}^{i} f(x_j)\, h_j \right| \le \frac{x_i - a}{x_N - a}\, TOL, \quad 1 \le i \le N, \tag{30.16}$$

i.e., the error grows at most linearly with x_i as i increases. However, this does not hold in general for the second adaptive algorithm. In Fig. 30.2, we plot the errors versus x_i for $x_i \le x_N$ resulting from the two adaptive algorithms with TOL = .01. We see the linear growth predicted for the first algorithm (30.12) while the error from the second algorithm (30.13) is larger for $1 < x_i < x_N$.

Chapter 30 Problems

30.1. Estimate the error using endpoint and midpoint quadrature for the following integrals: (a) $\int_0^2 2s\,ds$, (b) $\int_0^2 s^3\,ds$, and (c) $\int_0^2 \exp(-s)\,ds$ using $h = .1$, .01, .001 and .0001. Discuss the results.

30.2. Compute approximations of the following integrals using adaptive quadrature (a) $\int_0^2 2s\,ds$, (b) $\int_0^2 s^3\,ds$, and (c) $\int_0^2 \exp(-s)\,ds$. Discuss the results.

30.3. Compare theoretically and experimentally the number of steps of (30.12) and (30.13) for the computation of integrals of the form $\int_x^1 f(y)\,dy$ for $x > 0$, where $f(y) \sim y^{-\alpha}$ with $\alpha > 1$.

30.4. The *trapezoidal rule* takes the form

$$\int_{x_{j-1}}^{x_j} f(x)dx \approx (x_j - x_{j-1})(f(x_{j-1}) + f(x_j))/2. \tag{30.17}$$

Show that the quadrature is exact if $f(x)$ is a first order polynomial, and give an estimate of the quadrature error analogous to that of the midpoint rule. Compare the the midpoint and the trapezoidal method.

30.5. Design different adaptive quadrature algorithms based on the midpoint rule and make comparisons.

30.6. Consider a quadrature formula of the form

$$\int_a^b f(x)dx \approx (b-a)(f(\hat{x}_1) + f(\hat{x}_2))/2. \tag{30.18}$$

Determine the quadrature points $\hat{x}_1$ and $\hat{x}_2$, so that the quadrature formula is exact for $f(x)$ a second order polynomial. This quadrature rule is called the two-point Gauss rule. Check for which order of polynomials the resulting quadrature formula is exact.

30.7. Compute the value of $\int_0^1 \frac{1}{1+x^2}dx$ by quadrature. Multiply the result by 4. Do you recognize this number?

31

The exponential function $\exp(x) = e^x$

The need for mathematical skills is greater than ever, but it is widely recognized that, as a consequence of computer developments, there is a need for a shift in emphasis in the teaching of mathematics to students studying engineering. This shift is away from the simple mastery of solution techniques and towards development of a greater understanding of mathematical ideas and processes together with efficiency in applying this understanding to the formulation and analysis of physical phenomena and engineering systems. (Glyn James, in Preface to Modern Engineering Mathematics, 1992)

Because of the limitations of human imagination, one ought to say: everything is possible - and a bit more. (Horace Engdahl)

31.1 Introduction

In this chapter we return to study of the *exponential function* $\exp(x)$, which we have met above in Chapter *A very short course in Calculus* and Chapter *Galileo, Newton, Hooke, Malthus and Fourier*, and which is one of the basic functions of Calculus, see Fig. 35.5. We have said that $\exp(x)$ for $x > 0$ is the solution to the following initial value problem: Find a function $u(x)$ such that

$$\begin{aligned} u'(x) &= u(x) \quad \text{for } x > 0, \\ u(0) &= 1. \end{aligned} \tag{31.1}$$

Evidently, the equation $u'(x) = u(x)$ states that the rate of growth $u'(x)$ is equal to the quantity $u(x)$ itself, that is, the exponential function $\exp(x) =$

e^x is characterized by the property that its derivative is equal to itself: $D\exp(x) = \exp(x)$. What a wonderful almost divine property! We also denote the exponential function by e^x, that is, $e^x = \exp(x)$ and $De^x = e^x$.

In this chapter, we give a constructive proof of the *existence* of a unique solution to the initial value problem (31.1), that is, we *prove* the existence of the exponential function $\exp(x) = e^x$ for $x > 0$. Note that above, we just *claimed* the existence of solutions. As always, a constructive proof also shows how we may *compute* $\exp(x)$ for different values of x.

Below we extend $\exp(x)$ to $x < 0$ by setting $\exp(x) = (\exp(-x))^{-1}$ for $x < 0$, and show that $\exp(-x)$ solves the initial value problem $u'(x) = -u(x)$ for $x > 0$, $u(0) = 1$. We plot the functions $\exp(x)$ and $\exp(-x)$ for $x \geq 0$ in Fig. 35.5. We notice that $\exp(x)$ is increasing and $\exp(-x)$ is decreasing with increasing x, and that $\exp(x)$ is positive for all x. Combining $\exp(x)$ and $\exp(-x)$ for $x \geq 0$ defines $\exp(x)$ for $-\infty < x < \infty$. Below we show that $D\exp(x) = \exp(x)$ for $-\infty < x < \infty$.

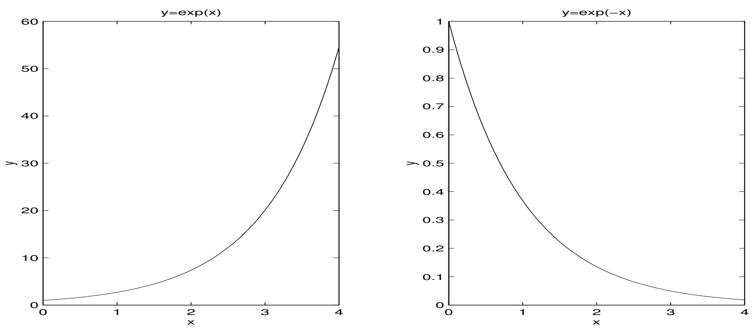


FIGURE 31.1. The exponential functions $\exp(x)$ and $\exp(-x)$ for $x \geq 0$.

The problem (31.1) is a special case of the Malthus population model (26.17), which also models a large variety of phenomena in e.g. physics and economy:

$$\begin{cases} u'(x) = \lambda u(x) & \text{for } x > 0, \\ u(0) = u_0. \end{cases} \tag{31.2}$$

where λ is a constant and u_0 is a given initial value. The solution of this problem can be expressed in terms of the exponential function as

$$u(x) = \exp(\lambda x)u_0 \quad \text{for } x \geq 0. \tag{31.3}$$

This follows directly from the fact that by the Chain rule, $D\exp(\lambda x) = \exp(\lambda x)\lambda$, where we used that $D\exp(x) = \exp(x)$. Assuming $u_0 > 0$ so that $u(x) > 0$, evidently the sign of λ determines if u decreases ($\lambda < 0$) or increases ($\lambda > 0$). In Fig. 35.5, we plotted the solutions of (31.2) with $\lambda = \pm 1$ and $u_0 = 1$.

Before going into the construction of the exponential function $\exp(x)$, we recall two of the key applications of (31.2): population dynamics and banking. Here x represents time and we change notation, replacing x by t.

EXAMPLE 31.1. We consider a population with constant birth and death rates β and δ, which are the numbers of births and deaths per individual creature per unit time. With $u(t)$ denoting the population at time t, there will be during the time interval from t to $t + \Delta t$ with Δt a small increment, approximately $\beta u(t)\Delta t$ births and $\delta u(t)\Delta t$ deaths. Hence the change in population over the time interval is approximately

$$u(t + \Delta t) - u(t) \approx \beta u(t)\Delta t - \delta u(t)\Delta t$$

and therefore

$$\frac{u(t + \Delta t) - u(t)}{\Delta t} \approx (\beta - \delta)u(t),$$

where the approximation improves as we decrease Δt. Taking the limit as $\Delta t \to 0$, assuming $u(t)$ is a differentiable function, we obtain the model $u'(t) = (\beta - \delta)u(t)$. Assuming the initial population at $t = 0$ is equal to u_0, leads to the model (31.2) with $\lambda = \beta - \delta$, with solution $u(x) = \exp(\lambda x)u_0$.

EXAMPLE 31.2. An investment u in a saving account earning 5% interest compounded continuously and beginning with \$2000 at time $t = 0$, satisfies

$$\begin{cases} u' = 1.05u, & t > 0, \\ u(0) = 2000, \end{cases}$$

and thus $u(t) = \exp(1.05t)2000$ for $t \geq 0$.

31.2 Construction of the exponential function $\exp(x)$ for $x \geq 0$

In the proof of the Fundamental Theorem, we constructed the solution $u(x)$ of the initial value problem

$$\begin{cases} u'(x) = f(u(x), x) & \text{for } 0 < x \leq 1, \\ u(0) = u_0, \end{cases} \tag{31.4}$$

in the case that $f(u(x), x) = f(x)$ depends only on x and not on $u(x)$. We constructed the solution $u(x)$ as the limit of a sequence of functions $\{U^n(x)\}_{n=1}^{\infty}$, where $U^n(x)$ is a piecewise linear function defined at a set of nodes $x_i^n = ih_n$, $i = 0, 1, 2, ..., N = 2^n$, $h_n = 2^{-n}$, by the relations

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n f(x_{i-1}^n) \quad \text{for } i = 1, 2, ..., N, \quad U^n(0) = u_0. \tag{31.5}$$

We shall now apply the same technique to construct the solution of (31.1), which has the form (31.4) with $f(u(x), x) = u(x)$ and $u_0 = 1$. We carry out the proof in a form which generalizes in a straight-forward way to any system of equations of the form (31.4), which really includes a very wide range of problems. We hope this will motivate the reader to carefully follow every step of the proof, to get properly prepared for the highlight Chapter *The general initial value problem.*

We construct the solution $u(x)$ of (31.1) for $x \in [0,1]$ as the limit of a sequence of piecewise linear functions $\{U^n(x)\}_{n=1}^{\infty}$ defined at the nodes by the formula

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n U^n(x_{i-1}^n) \quad \text{for } i = 1, 2, ..., N, \tag{31.6}$$

with $U^n(0) = 1$, which is an analog of (31.5) obtained by replacing $f(x_{i-1}^n)$ by $U^n(x_{i-1}^n)$ corresponding to replacing $f(x)$ by $f(x, u(x)) = u(x)$. Using the formula we can compute the values $U^n(x_i^n)$ one after the other for $i = 1, 2, 3, ...,$ starting from the initial value $U^n(0) = 1$, that is marching forward in time with x representing time.

We can write (31.6) in the form

$$U^n(x_i^n) = (1 + h_n)U^n(x_{i-1}^n) \quad \text{for } i = 1, 2, ..., N, \tag{31.7}$$

and conclude since $U^n(x_i^n) = (1 + h_n)U^n(x_{i-1}^n) = (1 + h_n)^2 U^n(x_{i-2}^n) = (1 + h_n)^3 U^n(x_{i-3}^n)$ and so on, that the nodal values of $U^n(x)$ are given by the formula

$$U^n(x_i^n) = (1 + h_n)^i, \quad \text{for } i = 0, 1, 2, ..., N, \tag{31.8}$$

where we also used that $U^n(0) = 1$. We illustrate in Fig. 31.2. We may

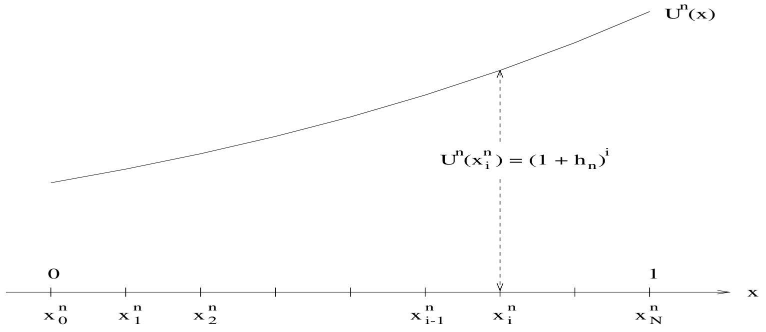


FIGURE 31.2. The piecewice linear approximate solution $U^n(x) = (1 + h_n)^i$.

view $U^n(x_i^n)$ as the capital obtained at time $x_i^n = ih_n$ starting with a unit capital at time zero, if the interest rate at each capitalization is equal to h_n.

To analyze the convergence of $U^n(x)$ as $n \to \infty$, we first prove a bound on the nodal values $U^n(x_i^n)$, by taking the logarithm of (31.8) and using the inequality $\log(1+x) \leq x$ for $x > 0$ from Problem 29.4, to obtain

$$\log(U^n(x_i^n) = i \log(1 + h_n) \leq i h_n = x_i^n \leq 1 \quad \text{for } i = 1, 2, .., N.$$

It follows that

$$U^n(x_i^n) = (1 + h_n)^i \leq 4 \quad \text{for } i = 1, 2, .., N, \tag{31.9}$$

since $\log(4) > 1$ according to Problem 29.1, and $\log(x)$ is increasing. Since $U^n(x)$ is linear between the nodes, and obviously $U^n(x) \geq 1$, we find that $1 \leq U^n(x) \leq 4$ for all $x \in [0, 1]$.

We now show that $\{U^n(x)\}_{n=1}^\infty$ is a Cauchy sequence for each fixed $x \in [0, 1]$. To see this, we first estimate $|U^n(x) - U^{n+1}(x)|$ at the node points $x = x_i^n = i h_n = 2i h_{n+1} = x_{2i}^{n+1}$ for $i = 0, 1, ..., N$, see Fig. 76.1. Notice that $h_{n+1} = h_n/2$ so that two steps with mesh size h_{n+1} corresponds to one step with mesh size h_n. We start by subtracting

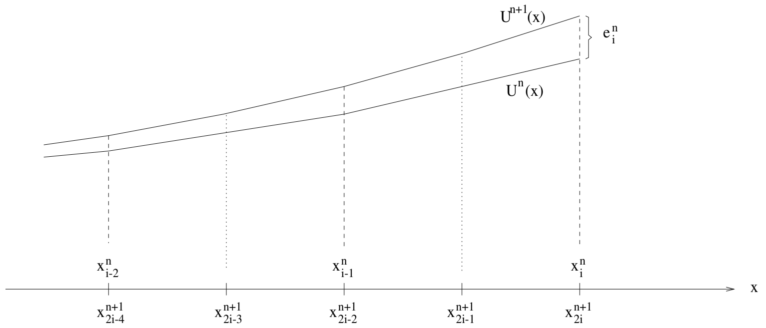


FIGURE 31.3. $U^n(x)$ and $U^{n+1}(x)$.

$$U^{n+1}(x_{2i}^{n+1}) = (1 + h_{n+1}) U^{n+1}(x_{2i-1}^{n+1}) = (1 + h_{n+1})^2 U^{n+1}(x_{2i-2}^{n+1}),$$

from (31.6), using that $x_i^n = x_{2i}^{n+1}$, and setting $e_i^n = U^n(x_i^n) - U^{n+1}(x_i^n)$, to get

$$e_i^n = (1 + h_n) U^n(x_{i-1}^n) - (1 + h_{n+1})^2 U^{n+1}(x_{i-1}^n),$$

which we may rewrite using that $(1 + h_{n+1})^2 = 1 + 2h_{n+1} + h_{n+1}^2$ and $2h_{n+1} = h_n$, as

$$e_i^n = (1 + h_n) e_{i-1}^n - h_{n+1}^2 U^{n+1}(x_{i-1}^n).$$

It follows using the bound $1 \leq U^{n+1}(x) \leq 4$ for $x \in [0, 1]$, that

$$|e_i^n| \leq (1 + h_n)|e_{i-1}^n| + 4h_{n+1}^2.$$

Inserting the corresponding estimate for e_{i-1}^n, we get

$$\begin{aligned}|e_i^n| &\leq (1+h_n)((1+h_n)|e_{i-2}^n| + 4h_{n+1}^2) + 4h_{n+1}^2 \\ &= (1+h_n)^2|e_{i-2}^n| + 4h_{n+1}^2(1+(1+h_n)).\end{aligned}$$

Continuing this way and using that $e_0^n = 0$, we obtain for $i = 1, ..., N$,

$$|e_i^n| \leq 4h_{n+1}^2 \sum_{k=0}^{i-1}(1+h_n)^k = h_n^2 \sum_{k=0}^{i-1}(1+h_n)^k.$$

Using now the fact that

$$\sum_{k=0}^{i-1} z^k = \frac{z^i - 1}{z-1} \tag{31.10}$$

with $z = 1 + h_n$, we thus obtain for $i = 1, ..., N$,

$$|e_i^n| \leq h_n^2 \frac{(1+h_n)^i - 1}{h_n} = h_n((1+h_n)^i - 1) \leq 3h_n,$$

where we again used that $(1+h_n)^i = U^n(x_i^n) \leq 4$. We have thus proved that for $\bar{x} = x_j^n$, $j = 1, ..., N$,

$$|U^n(\bar{x}) - U^{n+1}(\bar{x})| = |e_j^n| \leq 3h_n,$$

which is analogous to the central estimate (27.24) in the proof of the Fundamental Theorem.

Iterating this estimate over n as in the proof of (27.25), we get for $m > n$,

$$|U^n(\bar{x}) - U^m(\bar{x})| \leq 6h_n, \tag{31.11}$$

which shows that $\{U^n(\bar{x})\}_{n=1}^{\infty}$ is a Cauchy sequence and thus converges to a real number $u(\bar{x})$, which we choose to denote by $\exp(\bar{x}) = e^{\bar{x}}$. As in the proof of the Fundamental Theorem we can extend to a function $u(x) = \exp(x) = e^x$ defined for $x \in [0,1]$. Letting m tend to infinity in (31.11), we see that

$$|U^n(x) - \exp(x)| \leq 6h_n \quad \text{for } x \in [0,1]. \tag{31.12}$$

By the construction, we have if $\bar{x} = jh_n$ so that $h_n = \frac{\bar{x}}{j}$, noting that $j \to \infty$ as $n \to \infty$:

$$\exp(\bar{x}) = \lim_{n\to\infty}(1+h_n)^j = \lim_{j\to\infty}(1+\frac{\bar{x}}{j})^j,$$

that is,

$$\exp(x) = \lim_{j\to\infty}(1+\frac{x}{j})^j \qquad \text{for } x \in [0,1]. \tag{31.13}$$

In particular, we define the number e by

$$e \equiv \exp(1) = \lim_{j\to\infty}(1+\frac{1}{j})^j. \tag{31.14}$$

We refer to e as the *base of the exponential function*. We will prove below that $\log(e) = 1$.

It remains to verify that the function $u(x) = \exp(x) = e^x$ constructed above, indeed satifies (31.1) for $0 < x \leq 1$. We note that choosing $\bar{x} = jh_n$ and summing over i in (31.6), we get

$$U^n(\bar{x}) = \sum_{i=1}^{j} U^n(x_{i-1}^n)h_n + 1,$$

which we can write as

$$U^n(\bar{x}) = \sum_{i=1}^{j} u(x_{i-1}^n)h_n + 1 + E_n,$$

where $u(x) = \exp(x)$, and using (31.12),

$$|E_n| = |\sum_{i=1}^{j}(U^n(x_{i-1}^n) - u(x_{i-1}^n))h_n| \leq 6h_n \sum_{i=1}^{j} h_n \leq 6h_n,$$

since obviously $\sum_{i=1}^{j} h_n \leq 1$. Letting n tend to infinity and using $\lim_{n\to\infty} E_n = 0$, we see that $u(\bar{x}) = \exp(\bar{x})$ satisfies

$$u(\bar{x}) = \int_0^{\bar{x}} u(x)\,dx + 1.$$

Differentiating this equality with respect to $\bar{x}$, we get $u'(\bar{x}) = u(\bar{x})$ for $\bar{x} \in [0,1]$, and we have now proved that the constructed function $u(x)$ indeed solves the given initial value problem.

We conclude the proof by showing uniqueness. Thus, assume that we have two uniformly differentiable functions $u(x)$ and $v(x)$ such that $u'(x) = u(x)$ and $v'(x) = v(x)$ for $x \in (0,1]$, and $u(0) = v(0) = 1$. The $w = u - v$ satisfies $w'(x) = w(x)$ and $w(0) = 0$, and thus by the Fundamental Theorem,

$$w(x) = \int_0^x w'(y)\,dy = \int_0^x w(y)\,dy \quad \text{for } x \in [0,1].$$

Setting $a = \max_{0\leq x\leq 0.5}|w(x)|$, we thus have

$$a \leq \int_0^{0.5} a\,dy = 0.5a$$

which is possible only if $a = 0$ showing uniqueness for $0 \leq x \leq 0.5$. Repeating the argument on $[0.5, 1]$ proves that $w(x) = 0$ for $x \in [0, 1]$ and the uniqueness follows.

The proof immediately generalizes to $x \in [0, b]$ where b is any positive real number. We now summarize:

Theorem 31.1 *The initial value problem $u'(x) = u(x)$ for $x > 0$, and $u(0) = 1$, has a unique solution $u(x) = \exp(x)$ given by* (31.13).

31.3 Construction of $\exp(-x)$ for $x \geq 0$

If we define

$$\exp(-x) = \frac{1}{\exp(x)} \quad \text{for } x \geq 0,$$

then we find that

$$D\exp(-x) = D\frac{1}{\exp(x)} = -\frac{D\exp(x)}{(\exp(x))^2} = -\frac{1}{\exp(x)} = -\exp(-x). \quad (31.15)$$

We conclude that $\exp(-x)$ solves the initial value problem

$$u'(x) = -u(x) \quad \text{for } x > 0,\ u(0) = 1.$$

31.4 The exponential $\exp(x)$ for $x \in \mathbb{R}$

Piecing together the functions $\exp(x)$ and $\exp(-x)$ with $x \geq 0$, we obtain the function $u(x) = \exp(x)$ defined for $x \in \mathbb{R}$, which satisfies $u'(x) = u(x)$ for $x \in \mathbb{R}$ and $u(0) = 1$, see Fig. 31.4 and Fig. 31.5.

To see that $\frac{d}{dx}\exp(x)$ for $x < 0$, we set $y = -x > 0$ and compute $\frac{d}{dx}\exp(x) = \frac{d}{dy}\exp(-y)\frac{dy}{dx} = -\exp(-y)(-1) = \exp(x)$, where we used (31.15).

31.5 The basic property of the exponential: $\exp(a + b) = \exp(a)\exp(b)$

We now prove the basic property of the exponential function $\exp(x)$ using the fact that $\exp(x)$ satisfies the differential equation $D\exp(x) = \exp(x)$. We start considering the initial value problem

$$u'(x) = u(x) \quad \text{for } x > a, \quad u(a) = u_a, \quad (31.16)$$

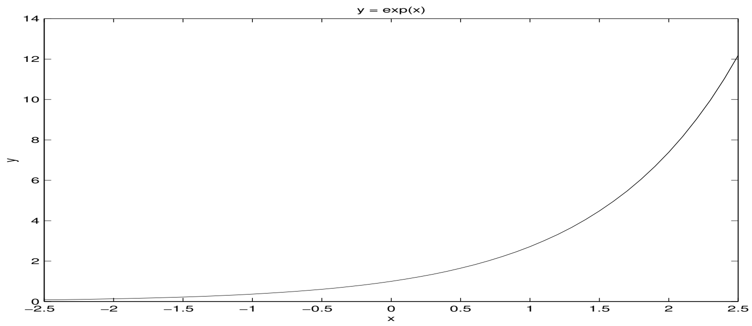


FIGURE 31.4. The exponential $\exp(x)$ for $x \in [-2.5, 2.5]$

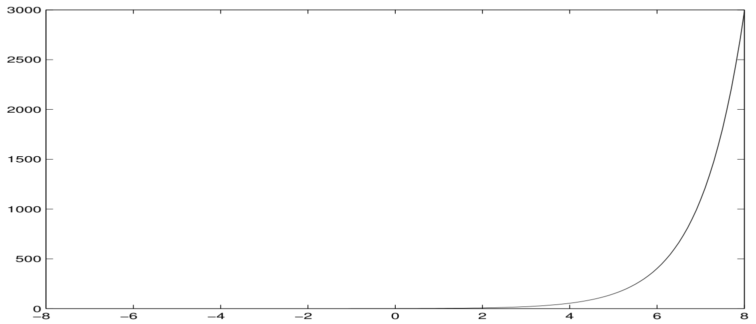


FIGURE 31.5. The exponential $\exp(x)$ for $x \in [-8, 8]$

with initial value at some point a other than zero. Setting $x = y + a$ and $v(y) = u(y+a) = u(x)$, we obtain by the Chain rule

$$v'(y) = \frac{d}{dy}u(y+a) = u'(y+a)\frac{d}{dy}(y+a) = u'(x),$$

and thus $v(y)$ satisfies the differential equation

$$v'(y) = v(y) \quad \text{for } y > 0, \quad v(0) = u_a.$$

This means that

$$v(y) = \exp(y)u_a \quad \text{for } y > 0.$$

Going back to the original variables, using that $y = x - a$, we find that the solution of (31.16) is given by

$$u(x) = \exp(x-a)u_a \quad \text{for } x \geq a. \tag{31.17}$$

We now prove that for $a, b \in \mathbb{R}$,

$$\exp(a+b) = \exp(a)\exp(b) \quad \text{or } e^{a+b} = e^a e^b, \tag{31.18}$$

which is the basic property of the exponential function. We do this by using the fact that $u(x) = \exp(x)$ satisfies the differential equation $u'(x) = u(x)$ and $\exp(0) = 1$. We have on the one hand that $u(a+b) = \exp(a+b)$ is the value of the solution $u(x)$ for $x = a+b$. We may reach to $x = a+b$, assuming $0 < a, b$ to start with, by first computing the solution $u(x) = \exp(x)$ from $x = 0$ up to $x = a$, which gives $u(a) = \exp(a)$. We next consider the following problem

$$v'(x) = v(x) \quad \text{for } x > a, \quad v(a) = \exp(a)$$

with solution $v(x) = \exp(x-a)\exp(a)$ for $x \geq a$. We have $v(x) = u(x)$ for $x \geq a$, since $u(x)$ also solves $u'(x) = u(x)$ for $x > a$, and $u(a) = \exp(a)$. Thus $v(b+a) = u(a+b)$, which translates into the desired equality $\exp(b)\exp(a) = \exp(a+b)$. The proof extends to any $a, b \in \mathbb{R}$.

31.6 The inverse of the exponential is the logarithm

We shall now prove that

$$\log(\exp(x)) = x \quad \text{for } x \in \mathbb{R}, \tag{31.19}$$

and conclude that

$$y = \exp(x) \quad \text{if and only if } x = \log(y), \tag{31.20}$$

which states that the inverse of the exponential is the logarithm.

We prove (31.19) by differentiation to get by the Chain rule for $x \in \mathbb{R}$,

$$\frac{d}{dx}(\log(\exp(x)) = \frac{1}{\exp(x)}\frac{d}{dx}(\exp(x)) = \frac{1}{\exp(x)}\exp(x) = 1,$$

and noting that $\log(\exp(0)) = \log(1) = 0$, which gives (31.19). Setting $x = \log(y)$ in (31.19), we have $\log(\exp(\log(y))) = \log(y)$, that is

$$\exp(\log(y)) = y \quad \text{for } y > 0. \tag{31.21}$$

We note in particular that

$$\exp(0) = 1 \text{ and } \log(e) = 1 \tag{31.22}$$

since $0 = \log(1)$ and $e = \exp(1)$ respectively.

In many Calculus books the exponential function $\exp(x)$ is defined as the inverse of the logarithm $\log(x)$ (which is defined as an integral). However, we prefer to directly prove the existence of $\exp(x)$ through its defining initial value problem, since this prepares the construction of solutions to general intial value problems.

31.7 The function a^x with $a > 0$ and $x \in \mathbb{R}$

We recall that in Chapter *The function* $y = x^r$ we defined the function x^r for $r = p/q$ rational with p and $q \neq 0$ integers, and x is a positive real number, as the solution y to the equation $y^q = x^p$.

We thus are familiar with a^x with $a > 0$ and x rational, and we may extend to $x \in \mathbb{R}$ by defining:

$$a^x = \exp(x \log(a)). \tag{31.23}$$

We now prove the basic properties of a^x with $x \in \mathbb{R}$, that is, the positive number a raised to the power $x \in \mathbb{R}$, extending our previous experience with x rational. We note that by the Chain rule the function $u(x) = a^x$ satisfies the differential equation

$$u'(x) = \log(a)u(x)$$

and $u(0) = 1$. In particular, choosing $a = e = \exp(1)$, we find that $a^x = e^x = \exp(x)$, and we thus conclude that the exponential function $\exp(x)$ indeed equals the number e raised to the power x. Note that before we just used e^x just as a another notation for $\exp(x)$.

Using now the exponential law (31.18) for $\exp(x)$, we obtain with a direct computation using the definition (31.23) the following analog for a^x:

$$a^{x+y} = a^x a^y. \tag{31.24}$$

The other basic rule for a^x reads:

$$(a^x)^y = a^{xy}, \tag{31.25}$$

which follows from the following computation:

$$(a^x)^y = \exp(y \log(a^x)) = \exp(y \log(\exp(x \log(a)))) = \exp(yx \log(a)) = a^{xy}.$$

As indicated, the rules (31.24) and (31.25) generalize the corresponding rules with x and y rational proved above.

We conclude computing the derivative of the function a^x from the definition (31.23) using the Chain rule:

$$\frac{d}{dx}a^x = \log(a)a^x. \tag{31.26}$$

Chapter 31 Problems

31.1. Define $U^n(x_i^n)$ alternatively by $U^n(x_i^n) = U^n(x_{i-1}^n) \pm h_n U^n(x_i^n)$, and prove that the corresponding sequence $\{U^n(x)\}$ converges to $\exp(\pm x)$.

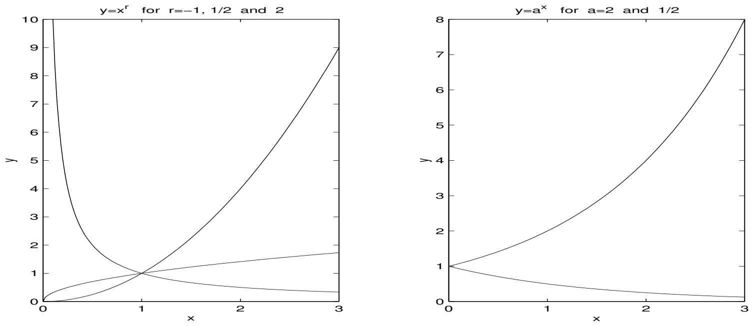


FIGURE 31.6. Examples of functions x^r and a^x.

31.2. Prove that for $x > 0$

$$(1 + \frac{x}{n})^n < \exp(x) \quad \text{for } n = 1, 2, 3 \ldots. \tag{31.27}$$

Hint: Take logarithm and use that $\log(1+x) < x$ for $x > 0$, and that the logarithm is increasing.

31.3. Prove directly the existence of a unique solution of $u'(x) = -u(x)$ for $x > 0$, $u(0) = 1$, that is construct $\exp(-x)$ for $x \geq 0$.

31.4. Show that the more often the bank capitalizes your interest, the better off you are, that is verify that

$$(1 + \frac{a}{n})^n \leq (1 + \frac{a}{n+1})^{n+1}. \tag{31.28}$$

Hint: Use the Binomial theorem.

31.5. Assume a bank offers "continuous capitalization" of the interest, corresponding to the (annual) interest rate a. What is then the "effective annual interest rate"?

31.6. Prove the differentiation formula $\frac{d}{dx}x^r = rx^{r-1}$ for $r \in \mathbb{R}$.

31.7. Prove the basic properties of the exponential function using that it is the inverse of the logarithm and use properties of the logarithm.

31.8. Given that the equation $u'(x) = u(x)$ has a solution for $x \in [0, 1]$ with $u(0) = 1$, construct a solution for all $x \geq 0$. Hint: use that if $u(x)$ satisfies $u'(x) = u(x)$ for $0 < x \leq 1$, then $v(x) = u(x - 1)$ satisfies $v'(x) = v(x)$ for $1 < x \leq 2$ and $v(1) = u(0)$.

31.9. Give the Taylor polynomial of order n with error term for $\exp(x)$ at $x = 0$.

31.10. Find a primitive function to (a) $x\exp(-x^2)$, (b) $x^3\exp(-x^2)$.

31.11. Compute the derivatives of the following functions: a) $f(x) = a^x$, $a > 0$, b) $f(x) = \exp(x+1)$, c) $f(x) = x\exp(x^2)$, d) $f(x) = x^3\exp(x^2)$, e) $f(x) = \exp(-x^2)$.

31.12. Compute the integrals $\int_0^1 f(x)dx$ of the functions in the previous exercise, except for the one in e), $f(x) = \exp(-x^2)$. Why do you think we left this one out?

31.13. Try to find the value of $\int_{-\infty}^{\infty} \exp(-x^2)dx$ numerically by quadrature. Square the result. Do you recognize this number?

31.14. Show that $\exp(x) \geq 1 + x$ for all x, not just for $x > -1$.

31.15. Show, by induction, that

$$\frac{d^n}{dx^n}\left(\mathrm{e}^x f(x)\right) = \mathrm{e}^x\left(1 + \frac{d}{dx}\right)^n f(x).$$

31.16. Prove (31.24) using the basic property (31.18) of the exponential and the definition (31.23).

31.17. Construct directly, without using the exponential function, the solution to the initial value problem $u'(x) = au(x)$ for $x \geq 0$ with $u(0) = 1$, where a is a real constant. Call the solution $\mathrm{aexp}(x)$. Prove that the function $\mathrm{aexp}(x)$ satisfies $\mathrm{aexp}(x+y) = \mathrm{aexp}(x)\mathrm{aexp}(y)$ for $x, y \geq 0$.

31.18. Define with $a > 0$ given, the function $y = \log_a(x)$ for $x > 0$ as the solution y to the equation $a^y = x$. With $a = e$ we get $\log_e(x) = \log(x)$, the natural logarithm. With $a = 10$ we get the so-called 10-logarithm. Prove that (i) $\log_a(xy) = \log_a(x) + \log_a(y)$ for $x, y > 0$, (ii) $\log_a(x^r) = r\log_a(x)$ for $x > 0$ and $r \in \mathbb{R}$, and (iii) $\log_a(x)\log(a) = \log(x)$ for $x > 0$.

31.19. Give the details of the proof of (31.26).

32
Trigonometric Functions

When I get to the bottom, I go back to the top of the slide where I stop and I turn and I go for a ride 'til I get to the bottom and I see you again. (Helter Skelter, Lennon-McCartney, 1968)

32.1 The defining differential equation

In this chapter, we shall study the following *initial value problem for a second order differential equation*: Find a function $u(x)$ defined for $x \geq 0$ satisfying

$$u''(x) = -u(x) \quad \text{for } x > 0, \quad u(0) = u_0,\ u'(0) = u_1, \tag{32.1}$$

where u_0 and u_1 are given *initial data*. We here require two initial conditions because the problem involves a second order derivative. We may compare with the first order initial value problem: $u'(x) = -u(x)$ for $x > 0, \quad u(0) = u_0$, with the solution $u(x) = \exp(-x)$, which we studied in the previous chapter.

We shall demonstrate below, in Chapter *The general initial value problem*, that (32.1) has a unique solution for any given values of u_0 and u_1, and we shall in this chapter show that the solution with initial data $u_0 = 0$ and $u_1 = 1$ is an old friend, namely, $u(x) = \sin(x)$, and the solution with $u_0 = 1$ and $u_1 = 0$ is $u(x) = \cos(x)$. Here $\sin(x)$ and $\cos(x)$ are the usual trigonometric functions defined geometrically in Chapter *Pythagoras and Euclid*, with the change that we measure the angle x in the unit of *radians* instead of degrees, with one radian being equal to $\frac{180}{\pi}$ degrees. In particular, we shall explain why one radian equals $\frac{180}{\pi}$ degrees.

We may thus define the trigonometric functions $\sin(x)$ and $\cos(x)$ as the solutions of (32.1) with certain initial data if we measure angles in the unit of radian. This opens a fresh route to understanding properties of the trigonometric functions by studying properties of solutions the differential equation (32.1), and we shall now explore this possibility.

We start by rewriting (32.1) changing the independent variable from x to t, since to aid our intuition we will use a mechanical interpretation of (32.1), where the independent variable represents time. We denote the derivative with respect to t with a dot, so that $\dot{u} = \frac{du}{dt}$, and $\ddot{u} = \frac{d^2u}{dt^2}$. We thus rewrite (32.1) as

$$\ddot{u}(t) = -u(t) \quad \text{for } t > 0, \quad u(0) = 0, \; \dot{u}(0) = 1, \tag{32.2}$$

where we chose $u_0 = 0$ and $u_1 = 1$ anticipating that we are looking for $\sin(t)$.

We now recall from Chapter *Galileo, Newton, Hooke, Malthus and Fourier*, that (32.2) is a model of the motion of unit mass along a friction-less horisontal x-axis with the mass connected to one end of a Hookean spring with spring constant equal to 1 and with the other end connected to the origin, see Fig. 26.3. We let $u(t)$ denotes the position ($x-$coordinate) of the mass at time t, and we assume that the mass is started at time $t = 0$ at the origin with speed $\dot{u}(0) = 1$, that is, $u_0 = 0$ and $u_1 = 1$. The spring exerts a force on the mass directed towards the origin, which is proportional to the length of the spring, since the spring constant is equal to 1, and the equation (32.2) expresses Newton's law: the acceleration $\ddot{u}(t)$ is equal to the spring force $-u(t)$. Because there is no friction, we would expect the mass to oscillate back and forth across the equilibrium position at the origin. We plot the solution $u(t)$ to (32.2) in Fig. 32.1, which clearly resembles the plot of the $\sin(t)$ function.

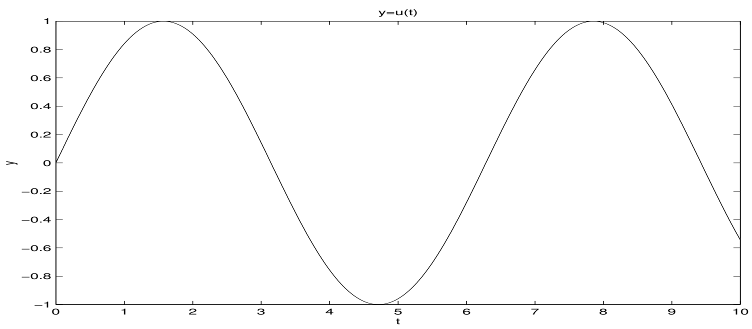


FIGURE 32.1. The solution of (32.2). Is it the function $\sin(t)$?

Let's now prove that our intuitive feeling indeed is correct, that is, let us prove that the solution of (32.2) indeed is our old friend $\sin(t)$. The key step then turns out to be to multiply our equation $\ddot{u} + u = 0$ by $\dot{u}$, to get

$$\frac{d}{dt}(\dot{u}^2 + u^2) = 2\dot{u}\ddot{u} + 2u\dot{u} = 2\dot{u}(\ddot{u} + u) = 0.$$

We conclude that $\dot{u}^2(t)+u^2(t)$ is constant for all t, and since $\dot{u}^2(0)+u^2(0) = 1 + 0 = 1$, we have found that the solution $u(t)$ of (32.2) satisfies the *conservation property*

$$\dot{u}^2(t) + u^2(t) = 1 \quad \text{for } t > 0, \tag{32.3}$$

which states that the point $(\dot{u}(t), u(t)) \in \mathbb{R}^2$ lies on the unit circle in $\mathbb{R}^2$, see Fig. 32.2.

We remark that in mechanical terms, the relation (32.3) expresses that the *total energy*

$$E(t) \equiv \frac{1}{2}\dot{u}^2(t)) + \frac{1}{2}u^2(t), \tag{32.4}$$

is preserved ($= 1/2$) during the motion. The total energy at time t is the sum of the *kinetic energy* $\dot{u}^2(t)/2$, and the *potential energy* $u^2(t)/2$. The potential energy is the energy stored in the spring, which is equal to the *work* $W(u(t))$ to stretch the spring the distance $u(t)$:

$$W(u(t)) = \int_0^{u(t)} v\, dv = \frac{1}{2}u^2(t),$$

where we used the principle that to stretch the spring from v to $v + \Delta v$, the work is $v\Delta v$ since the spring force is v. At the extreme points with $\dot{u}(t) = 0$, the kinetic energy is zero and all energy occurs as potential energy, while all energy occurs as kinetic energy when the body passes the origin with $u(t) = 0$. During the motion of the body, the energy is thus periodiaclly transferred from kinetic energy to potential energy an back again.

Going now back to (32.3), we thus see that the point $(\dot{u}(t), u(t)) \in \mathbb{R}^2$ moves on the unit circle and the *velocity* of the motion is given by $(\ddot{u}(t), \dot{u}(t))$, which we obtain by differentiating each coordinate function with respect to t. We will return to this issue in Chapter *Curves* below. Using the differential equation $\ddot{u} + u = 0$, we see that

$$(\ddot{u}(t), \dot{u}(t)) = (-u(t), \dot{u}(t)),$$

and conclude recalling (32.3) that the modulus of the velocity is equal to 1 for all t. We conclude that the point $(\dot{u}(t), u(t))$ moves around the unit circle with unit velocity and at time $t = 0$ the point is at position $(1, 0)$. But this directly connects with the usual geometrical definition of $(\cos(t), \sin(t))$ as the coordinates of a point on the unit circle at the angle t, see Fig. 32.2, so that we should have $(\dot{u}(t), u(t)) = (\cos(t), \sin(t))$. To make this connection straight, we of course need to measure angles properly, and the proper measure is *radians* with 2π radians corresponding to 360 degrees. This is because the time for one revolution with speed 1 should be equal to 2π, that is the length of the circumference of the unit circle.

In fact, we can use the solution $\sin(t)$ of the initial value problem (32.2) to define the number π as the smallest positive root $\bar{t}$ of $\sin(t)$, corresponding to one half revolution with $u(\bar{t}) = 0$ and $\dot{u}(\bar{t}) = -1$.

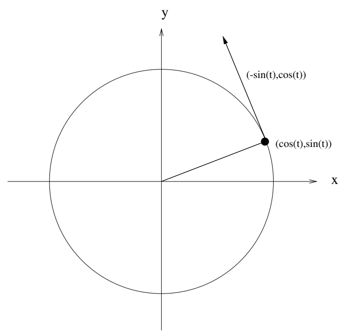


FIGURE 32.2. Energy conservation.

We may now conclude that the solution $u(t)$ of (32.2) satisfies $(\dot{u}(t), u(t)) = (\cos(t), \sin(t))$, so that in particular $u(t) = \sin(t)$ and $\frac{d}{dt}\sin(t) = \cos(t)$, where $(\cos(t), \sin(t))$ is defined geometrically as the point on the unit circle of angle t radians.

We can now turn the argument around, and simply define $\sin(t)$ as the solution $u(t)$ to (32.2) with $u_0 = 0$ and $u_1 = 1$, and then define $\cos(t) = \frac{d}{dt}\sin(t)$. Alternatively, we can define $\cos(t)$ as the solution of the problem

$$\ddot{v}(t) = -v(t) \quad \text{for } t > 0, \quad v(0) = 1,\ \dot{v}(0) = 0, \tag{32.5}$$

which we obtain by differentiation of (32.2) with respect to t and using the initial conditions for $\sin(t)$. Differentating once more, we see that $\frac{d}{dt}\cos(t) = -\sin(t)$.

Both $\sin(t)$ and $\cos(t)$ will be *periodic with period* 2π, because the point $(\dot{u}(t), u(t))$ moves around the unit circle with velocity one and comes back the same point after a time period of 2π. As we said, we may in particular define π as the first value of $t > 0$ for which $\sin(t) = 0$, which corresponds the point $(\dot{u}, u) = (-1, 0)$, and 2π will then be time it takes for the point $(\dot{u}, u)$ to make one complete revolution starting at $(1, 0)$, moving to $(-1, 0)$ following the upper semi-circle and then returning to $(1, 0)$ following the lower semi-circle. The periodicity of $u(t)$ with period 2π is expressed as

$$u(t + 2n\pi) = u(t) \quad \text{for } t \in \mathbb{R},\ n = 0, \pm 1, \pm 2, ... \tag{32.6}$$

The energy conservation (32.3) translates into the most well known of all trigonometric formulas:

$$\sin^2(t) + \cos^2(t) = 1 \quad \text{for } t > 0. \tag{32.7}$$

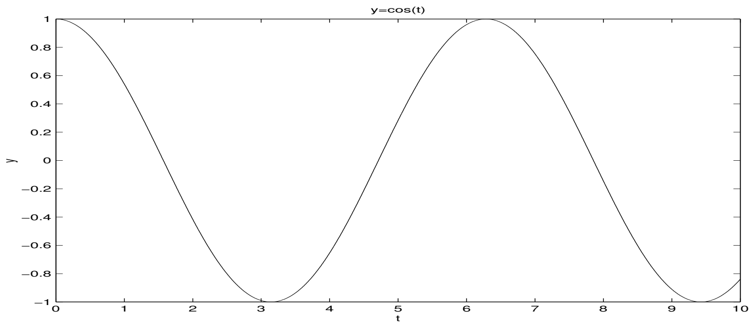


FIGURE 32.3. The function $\cos(t)$!

To compute the values of $\sin(t)$ and $\cos(t)$ for a given t, we may compute the solution to the corresponding defining differential initial value problem. We return to this topic below.

We summarize:

Theorem 32.1 *The initial value problem $u''(x) + u(x) = 0$ for $x > 0$ with $u_0 = 0$ and $u_1 = 1$, has a unique solution, which is denoted by $\sin(x)$. The initial value problem $u''(x) + u(x) = 0$ for $x > 0$ with $u_0 = 1$ and $u_1 = 0$, has a unique solution, which is denoted by $\cos(x)$. The functions $\sin(x)$ and $\cos(x)$ extend to $x < 0$ as solutions of $u''(x) + u(x) = 0$ and are periodic with period 2π, and $\sin(\pi) = 0$, $\cos(\frac{\pi}{2}) = 0$. We have $\frac{d}{dx}\sin(x) = \cos(x)$ and $\frac{d}{dx}\cos(x) = -\sin(x)$. Further $\cos(-x) = \cos(x)$, $\cos(\pi - x) = -\cos(x)$, $sin(\pi - x) = \sin(x)$, $sin(-x) = -\sin(x)$, $\cos(x) = \sin(\frac{\pi}{2} - x)$, $\sin(x) = \cos(\frac{\pi}{2} - x)$, $\sin(\frac{\pi}{2} + x) = \cos(x)$, and $\cos(\frac{\pi}{2} + x) = -\sin(x)$.*

32.2 Trigonometric formulas

Using the defining differential equation $u''(x) + u(x) = 0$, we can verify the following basic trigonometric formulas for $x, y \in \mathbb{R}$:

$$\sin(x + y) = \sin(x)\cos(y) + \cos(x)\sin(y) \tag{32.8}$$

$$\sin(x - y) = \sin(x)\cos(y) - \cos(x)\sin(y) \tag{32.9}$$

$$\cos(x + y) = \cos(x)\cos(y) - \sin(x)\sin(y) \tag{32.10}$$

$$\cos(x - y) = \cos(x)\cos(y) + \sin(x)\sin(y). \tag{32.11}$$

For example, to prove (41.12), we note that both the right hand and left hand side satisfy the equation $u''(x) + u(x) = 0$, and the initial conditions $u(0) = \sin(y)$, $u'(0) = \cos(y)$, with y acting as a parameter, and thus are equal.

We note the particular special cases:

$$\sin(2x) = 2\sin(x)\cos(x) \tag{32.12}$$

$$\cos(2x) = \cos^2(x) - \sin^2(x) = 2\cos^2(x) - 1 = 1 - 2\sin^2(x). \tag{32.13}$$

Adding (41.12) and (41.12), we obtain

$$\sin(x + y) + \sin(x - y) = 2\sin(x)\cos(y).$$

Setting $\bar{x} = x + y$ and $\bar{y} = x - y$ we obtain the first of the following set of formulas, all proved similarly,

$$\sin(\bar{x}) + \sin(\bar{y}) = 2\sin(\frac{\bar{x}+\bar{y}}{2})\cos(\frac{\bar{x}-\bar{y}}{2}) \tag{32.14}$$

$$\sin(\bar{x}) - \sin(\bar{y}) = 2\cos(\frac{\bar{x}+\bar{y}}{2})\sin(\frac{\bar{x}-\bar{y}}{2}) \tag{32.15}$$

$$\cos(\bar{x}) + \cos(\bar{y}) = 2\cos(\frac{\bar{x}+\bar{y}}{2})\cos(\frac{\bar{x}-\bar{y}}{2}) \tag{32.16}$$

$$\cos(\bar{x}) - \cos(\bar{y}) = -2\sin(\frac{\bar{x}+\bar{y}}{2})\sin(\frac{\bar{x}-\bar{y}}{2}). \tag{32.17}$$

32.3 The functions $\tan(x)$ and $\cot(x)$ and their derivatives

We define

$$\tan(x) = \frac{\sin(x)}{\cos(x)}, \quad \cot(x) = \frac{\cos(x)}{\sin(x)}, \tag{32.18}$$

for values of x such that the denominator is different from zero. We compute the derivatives

$$\frac{d}{dx}\tan(x) = \frac{\cos(x)\cos(x) - \sin(x)(-\sin(x))}{\cos^2(x)} = \frac{1}{\cos^2(x)}, \tag{32.19}$$

and similarly

$$\frac{d}{dx}\cot(x) = -\frac{1}{\sin^2(x)}. \tag{32.20}$$

Dividing (41.12) by (41.12), and dividing both numerator and denominator by $\cos(x)\cos(y)$, we obtain

$$\tan(x + y) = \frac{\tan(x) + \tan(y)}{1 - \tan(x)\tan(y)}, \tag{32.21}$$

and similarly,

$$\tan(x-y) = \frac{\tan(x) - \tan(y)}{1 + \tan(x)\tan(y)}. \tag{32.22}$$

32.4 Inverses of trigonometric functions

Inverses of the basic trigonometric functions $\sin(x)$, $\cos(x)$, $\tan(x)$ and $\cot(x)$, are useful in applications. We now introduce and give names to these inverses and derive their basic properties.

The function $f(x) = \sin(x)$ is strictly increasing from -1 to 1 on $[\frac{-\pi}{2}, \frac{\pi}{2}]$, because the derivative $f'(x) = \cos(x)$ is positive on $(\frac{-\pi}{2}, \frac{\pi}{2})$. Thus the function $y = f(x) = \sin(x)$ with $D(f) = [\frac{-\pi}{2}, \frac{\pi}{2}]$ and $R(f) = [-1, 1]$, therefore has an inverse $x = f^{-1}(y)$, which we denote by

$$x = f^{-1}(y) = \arcsin(y), \tag{32.23}$$

and $D(f^{-1}) = D(\arcsin) = [-1, 1]$ and $R(f^{-1}) = R(\arcsin) = [\frac{-\pi}{2}, \frac{\pi}{2}]$, see Fig. 32.5.

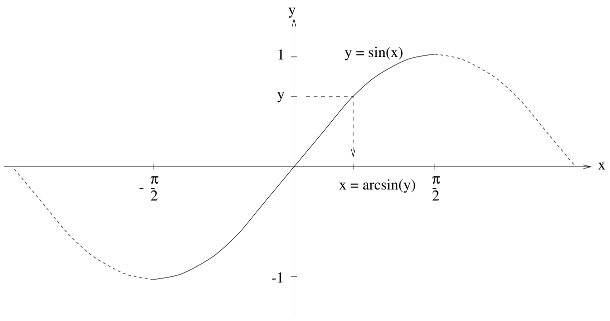


FIGURE 32.4. The function $x = \arcsin(y)$

We thus have

$$\sin(\arcsin(y)) = y \quad \arcsin(\sin(x)) = x \quad \text{for } x \in [\frac{-\pi}{2}, \frac{\pi}{2}], y \in [-1, 1]. \tag{32.24}$$

We next compute the derivative of $\arcsin(y)$ with respect to y:

$$\frac{d}{dy}\arcsin(y) = \frac{1}{\frac{d}{dx}\sin(x)} = \frac{1}{\cos(x)} = \frac{1}{\sqrt{1 - \sin^2(x)}} = \frac{1}{\sqrt{1 - y^2}}.$$

Similarly, the function $y = f(x) = \tan(x)$ is strictly increasing on $D(f) = (\frac{-\pi}{2}, \frac{\pi}{2})$ and $R(f) = \mathbb{R}$, and thus has an inverse, which we denote by

$$x = f^{-1}(y) = \arctan(y),$$

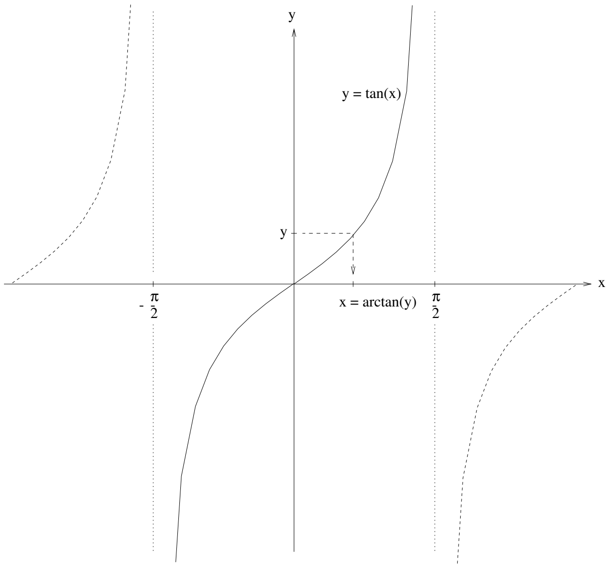


FIGURE 32.5. The function $x = \arctan(y)$

with $D(\arctan) = \mathbb{R}$ and $R(\arctan) = (\frac{-\pi}{2}, \frac{\pi}{2})$, see Fig. 34.3.

We compute the derivative of $\arctan(y)$:

$$\frac{d}{dy}\arctan(y) = \frac{1}{\frac{d}{dx}\tan(x)} = \cos^2(x)$$
$$= \frac{\cos^2(x)}{\cos^2(x)+\sin^2(x)} = \frac{1}{1+\tan^2(x)} = \frac{1}{1+y^2}.$$

We define similarly the inverse of $y = f(x) = \cos(x)$ with $D(f) = [0, \pi]$ and denote the inverse by $x = f^{-1}(y) = \arccos(y)$ with $D(\arccos) = [-1, 1]$ and $R(\arccos) = [0, \pi]$. We have

$$\frac{d}{dy}\arccos(y) = \frac{1}{\frac{d}{dx}\cos(x)} = -\frac{1}{\sin(x)} = -\frac{1}{\sqrt{1-\cos^2(x)}} = -\frac{1}{\sqrt{1-y^2}}.$$

Finally, we define the inverse of $y = f(x) = \cot(x)$ with $D(f) = (0, \pi)$ and denote the inverse by $x = f^{-1}(y) = \operatorname{arccot}(y)$ with $D(\operatorname{arccot}) = \mathbb{R}$ and $R(\operatorname{arccot}) = (0, \pi)$. We have

$$\frac{d}{dy}\operatorname{arccot}(y) = \frac{1}{\frac{d}{dx}\cot(x)} = -\sin^2(x) = -\frac{\sin^2(x)}{\cos^2(x)+\sin^2(x)}$$
$$= -\frac{1}{1+\cot^2(x)} = -\frac{1}{1+y^2}.$$

We summarize:

$$\begin{aligned}
\frac{d}{dx}\arcsin(x) &= \frac{1}{\sqrt{1-x^2}} \quad \text{for } x \in (-1,1)\\
\frac{d}{dx}\arccos(x) &= -\frac{1}{\sqrt{1-x^2}} \quad \text{for } x \in (-1,1)\\
\frac{d}{dx}\arctan(x) &= \frac{1}{1+x^2} \quad \text{for } x \in \mathbb{R}\\
\frac{d}{dx}\mathrm{arccot}(x) &= -\frac{1}{1+x^2} \quad \text{for } x \in \mathbb{R}.
\end{aligned} \tag{32.25}$$

In other words,

$$\begin{aligned}
\arcsin(x) &= \int_0^x \frac{1}{\sqrt{1-y^2}}\,dy \quad \text{for } x \in (-1,1)\\
\arccos(x) &= \frac{\pi}{2} - \int_0^x \frac{1}{\sqrt{1-y^2}}\,dy \quad \text{for } x \in (-1,1)\\
\arctan(x) &= \int_0^x \frac{1}{1+y^2}\,dy \quad \text{for } x \in \mathbb{R}\\
\mathrm{arccot}(x) &= \frac{\pi}{2} - \int_0^x \frac{1}{1+y^2}\,dy \quad \text{for } x \in \mathbb{R}.
\end{aligned} \tag{32.26}$$

We also note the following analog of (32.21) obtained by setting $x = \arctan(u)$ and $y = \arctan(v)$, so that $u = \tan(x)$ and $v = \tan(y)$, and assuming that $x + y \in (-\frac{\pi}{2}, \frac{\pi}{2})$:

$$\arctan(u) + \arctan(v) = \arctan(\frac{u+v}{1-uv}). \tag{32.27}$$

32.5 The functions sinh(x) and cosh(x)

We define for $x \in \mathbb{R}$

$$\sinh(x) = \frac{e^x - e^{-x}}{2} \quad \text{and} \quad \cosh(x) = \frac{e^x + e^{-x}}{2}. \tag{32.28}$$

We note that

$$D\sinh(x) = \cosh(x) \quad \text{and} \quad D\cosh(x) = \sinh(x). \tag{32.29}$$

We have $y = f(x) = \sinh(x)$ is strictly increasing and thus has an inverse $x = f^{-1}(y) = \mathrm{arcsinh}(y)$ with $D(\mathrm{arcsinh}) = \mathbb{R}$ and $R(\mathrm{arcsinh}) = \mathbb{R}$. Further, $y = f(x) = \cosh(x)$ is strictly increasing on $[0, \infty)$, and thus has an inverse $x = f^{-1}(y) = \mathrm{arccosh}(y)$ with $D(\mathrm{arccosh}) = [1, \infty)$ and $R(\mathrm{arccosh}) = [0, \infty)$. We have

$$\frac{d}{dy}\mathrm{arcsinh}(y) = \frac{1}{\sqrt{y^2+1}}, \quad \frac{d}{dy}\mathrm{arccosh}(y) = \frac{1}{\sqrt{y^2-1}}. \tag{32.30}$$

32.6 The hanging chain

Consider a hanging chain fixed at $(-1,0)$ and $(1,0)$ in a coordinate system with the x-axis horisontal and y-axis vertical. Let us seek the curve $y = y(x)$ described by the chain. Let $(F_h(x), F_v(x))$ be the two components of the *force* in the chain at x. Vertical and horisontal equilibrium of the element of the chain between x and $x + \Delta x$ gives

$$F_h(x + \Delta x) = F_h(x), \quad F_v(x) + m\Delta s = F_v(x + \Delta x),$$

where $\Delta s \approx \sqrt{(\Delta x)^2 + (\Delta y)^2} \approx \sqrt{1 + (y'(x))^2}\Delta x$, and m is the weight of the chain per unit length. We conclude that $F_h(x) = F_h$ is constant, and

$$F_v'(x) = m\sqrt{1 + (y'(x))^2}.$$

Momentum equilibrium around the midpoint of the element of the chain between x and $x + \Delta x$, gives

$$F_h \Delta y = \frac{1}{2}F_v(x + \Delta x)\Delta x + \frac{1}{2}F_v(x)\Delta x \approx F_v(x)\Delta x,$$

which leads to

$$y'(x) = \frac{F_v(x)}{F_h}. \tag{32.31}$$

Assuming that $F_h = 1$, we are thus led to the differential equation

$$F_v'(x) = m\sqrt{1 + (F_v(x))^2}.$$

We can check by direct differentation that this differential equation is satisfied if $F_v(x)$ solves the equation

$$\operatorname{arcsinh}(F_v(x)) = mx,$$

and we also have $F_v(0) = 0$. Therefore

$$F_v(x) = \sinh(mx),$$

and thus by (32.31),

$$y(x) = \frac{1}{m}\cosh(mx) + c$$

with the constant c to be chosen so that $y(\pm 1) = 0$. We thus obtain the following solution

$$y(x) = \frac{1}{m}(\cosh(mx) - \cosh(m)). \tag{32.32}$$

The curve $y(x) = \cosh(mx) + c$ with m and c constants, is called the *hanging chain curve*, or the *catenaria*.

32.7 Comparing solutions of $u'' + k^2u(x) = 0$ and $u'' - k^2u(x) = 0$

We summarize some experience from above. The solutions of the equation $u'' + k^2u(x) = 0$ are linear combinations of $\sin(kx)$ and $\cos(kx)$. The solutions of $u'' - k^2u(x) = 0$ are linear combinations of $\sinh(kx)$ and $\cosh(kx)$.

Chapter 32 Problems

32.1. Show that the solution of $\ddot{u}(t)+u(t) = 0 \quad \text{for } t > 0,\ u(0) = \sin(\alpha),\ u'(0) = \cos(\alpha)$, is given by $u(t) = \cos(t)\sin(\alpha) + \sin(t)\cos(\alpha) = \sin(t + \alpha)$.

32.2. Show that the solution of $\ddot{u}(t)+u(t) = 0 \quad \text{for } t > 0,\ u(0) = r\cos(\alpha),\ u'(0) = r\sin(\alpha)$, is given by $u(t) = r(\cos(t)\cos(\alpha) + \sin(t)\sin(\alpha)) = r\cos(t - \alpha)$.

32.3. Show that the solution to $\ddot{u}(t)+ku(t) = 0 \quad \text{for } t > 0,\ u(0) = r\cos(\alpha),\ u'(0) = r\sin(\alpha)$, where k is a given positive constant, is given by $r\cos(\sqrt{k}(t - \alpha))$. Give a mechanical interpretation of this model.

32.4. Show that the function $\sin(nx)$ solves the boundary value problem $u''(x)+n^2u(x) = 0$ for $0 < x < \pi$, $u(0) = u(\pi) = 0$.

32.5. Solve $u'(x) = \sin(x), \quad x > \pi/4,\ u(\pi/4) = 2/3$.

32.6. Show that (a) $\sin(x) < x$ for $x > 0$, (b) $x < \tan(x)$ for $0 < x < \frac{\pi}{2}$.

32.7. Show that $\lim_{x\to 0} \frac{\sin(x)}{x} = 1$.

32.8. Show the following relations from the definition, i.e. from the differential equation defining $\sin(x)$ and $\cos(x)$: (a) $\sin(-x) = -\sin(x)$, (b) $\cos(-x) = \cos(x)$, (c) $\sin(\pi - x) = \sin(x)$, (d) $\cos(\pi - x) = -\cos(x)$, (e) $\sin(\pi/2 - x) = \cos(x)$, (f) $\cos(\pi/2 - x) = \sin(x)$.

32.9. Prove the product formulas show that

$$\begin{aligned}
\sin(x)\sin(y) &= \frac{1}{2}\left(\cos(x-y) - \cos(x+y)\right),\\
\cos(x)\cos(y) &= \frac{1}{2}\left(\cos(x-y) + \cos(x+y)\right),\\
\sin(x)\cos(y) &= \frac{1}{2}\left(\sin(x-y) + \sin(x+y)\right).
\end{aligned}$$

32.10. Compute the following integrals by integrating by parts: (a) $\int_0^1 x^3\sin(x)dx$, (b) $\int_0^1 \exp(x)\sin(x)dx$, (c) $\int_0^1 x^2\cos(x)dx$.

32.11. Determine Taylor's formula for $\arctan(x)$ at $x = 0$ and use your result to calculate approximations of π. Hint: $\arctan(1) = \frac{\pi}{4}$.

32.12. Show that $\arctan(1) = \arctan(1/2) + \arctan(1/3)$. Try to find other rational numbers a and b such that $\arctan(1) = \arctan(a) + \arctan(b)$. In particular seek to find a and b as small as possible.

32.13. Combine your results from the previous two exercises to construct a better algorithm for computing π. Even more efficient methods may be obtained using the identity $\pi/4 = 4\arctan(1/5) - \arctan(1/239)$. Compare the two algorithms and explain why the second is more efficient.

32.14. Show that: (a) $\arcsin(-x) = -\arcsin(x)$, (b) $\arccos(-x) = \pi - \arccos(x)$, (c) $\arctan(-x) = -\arctan(x)$, (d) $\text{arccot}(-x) = \pi - \text{arccot}(x)$, (e) $\arcsin(x) + \arccos(x) = \pi/2$, (f) $\arctan(x) + \text{arccot}(x) = \pi/2$.

32.15. Calculate analytically: (a) $\arctan(\sqrt{2}-1)$, (b) $\arctan(1/8)+\arctan(7/9)$, (c) $\arcsin(1/7) + \arcsin(11/4)$, (d) $\tan(\arcsin(3/5)/2)$, (e) $\sin(2\arcsin(0.8))$, (f) $\arctan(2) + \arcsin(3/\sqrt{10})$.

32.16. Solve the equation: (a) $\arccos(2x) = \arctan(x)$, (b) $\arcsin(\cos(x)) = x\sqrt{3}$.

32.17. Calculate the derivative, if possible, of (a) $\arctan(\sqrt{x}-x^5)$, (b) $\arcsin(1/x^2)\arcsin(x^2)$, (c) $\tan(\arcsin(x^2))$, (d) $1/\arctan(\sqrt{x})$.

32.18. Compute numerically for different values of x, (a) $\arcsin(x)$, (b) $\arccos(x)$, (c) $\arctan(x)$, (d) $\text{arccot}(x)$.

32.19. Prove (32.30).

32.20. Verify that $\cosh^2(x) - \sinh^2(x) = 1$.

32.21. (a) Find the inverse $x = \text{arcsinh}(y)$ of $y = sinh(x) = \frac{1}{2}(e^x - e^{-x})$ by solving for x in terms of y. Hint: Multiply by e^x and solve for $z = e^x$. Then take logarithms. (b) Find a similar formula for $\text{arccosh}(y)$.

32.22. Compute analytically the area of a disc of radius 1 by computing the integral

$$\int_{-1}^{1} \sqrt{1-x^2}\,dx.$$

How do you handle the fact that $\sqrt{1-x^2}$ is not Lipschitz continuous on $[-1,1]$? Hint: Use the substitution $x = \sin(y)$ and the fact the $\cos^2(y) = \frac{1}{2}(1+\cos(2y))$.

33

The functions $\exp(z)$, $\log(z)$, $\sin(z)$ and $\cos(z)$ for $z \in \mathbb{C}$

The shortest path between two truths in the real domain passes through the complex domain. (Hadamard 1865-1963)

33.1 Introduction

In this chapter we extend some of the elementary functions to complex arguments. We recall that we can write a complex number z in the form $z = |z|(\cos(\theta) + i\sin(\theta))$ with $\theta = \arg z$ the argument of z, and $0 \leq \theta = \text{Arg } z < 2\pi$ the principal argument of z.

33.2 Definition of $\exp(z)$ for $z \in \mathbb{C}$

We define, writing $z = x + iy$ with $x, y \in \mathbb{R}$,

$$\exp(z) = e^z = e^x(\cos(y) + i\sin(y)), \tag{33.1}$$

which extends the definition of e^z with $z \in \mathbb{R}$ to $z \in \mathbb{C}$. We note that in particular for $y \in \mathbb{R}$,

$$e^{iy} = \cos(y) + i\sin(y), \tag{33.2}$$

which is also referred to as *Eulers formula*. We note that

$$\sin(y) = \frac{e^{iy} - e^{-iy}}{2i}, \quad \cos(y) = \frac{e^{iy} + e^{-iy}}{2}, \quad \text{for } y \in \mathbb{R}, \tag{33.3}$$

and

$$|e^{iy}| = 1 \quad \text{for } y \in \mathbb{R}. \tag{33.4}$$

We can now express a complex number $z = r(\cos(\theta) + i\sin(\theta))$ in the form

$$z = re^{i\theta} \tag{33.5}$$

with $\theta = \arg z$ and $r = |z|$.

One can prove (using the basic trigonometric formulas) that $\exp(z)$ satisfies the usual law for exponentials so that in particular for $z, \zeta \in \mathbb{C}$,

$$e^z e^\zeta = e^{z+\zeta}. \tag{33.6}$$

In particular, the rule for multiplication of two complex numbers $z = |z|e^{i\theta}$ and $\zeta = |\zeta|e^{i\varphi}$ can be expressed as follows:

$$z\zeta = |z|e^{i\theta}|\zeta|e^{i\varphi} = |z||\zeta|e^{i(\theta+\varphi)}. \tag{33.7}$$

33.3 Definition of sin(z) and cos(z) for $z \in \mathbb{C}$

We define for $z \in \mathbb{C}$

$$\sin(z) = \frac{e^{iz} - e^{-iz}}{2i}, \quad \cos(z) = \frac{e^{iz} + e^{-iz}}{2}, \tag{33.8}$$

which extends (33.3) to $\mathbb{C}$.

33.4 de Moivres formula

We have for $\theta \in \mathbb{R}$ and n an integer

$$(e^{i\theta})^n = e^{in\theta},$$

that is,

$$(\cos(\theta) + i\sin(\theta))^n = \cos(n\theta) + i\sin(n\theta), \tag{33.9}$$

which is referred to as *de Moivres formula*. In particular,

$$(\cos(\theta) + i\sin(\theta))^2 = \cos(2\theta) + i\sin(2\theta),$$

from which follows separating into real and complex parts

$$\cos(2\theta) = \cos^2(\theta) - \sin^2(\theta), \quad \sin(2\theta) = 2\cos(\theta)\sin(\theta).$$

Using de Moivres formula gives a quick way of deriving some of the basic trigonometric formulas (in case one has forgotten these formulas).

33.5 Definition of $\log(z)$

We have defined above $\log(x)$ for $x > 0$ and we now pose the problem of defining $\log(z)$ for $z \in \mathbb{C}$. We recall that $w = \log(x)$ can be viewed as the unique solution to the equation $e^w = x$, where $x > 0$. We consider therefore the equation

$$e^w = z,$$

with $z = |z|(\cos(\theta) + i\sin(\theta)) \in \mathbb{C}$ being given assuming $z \neq 0$, and we seek $w = \mathrm{Re}\ w + i\mathrm{Im}\ w \in \mathbb{C}$, with the intention to call a solution $w = \log(z)$. Here Re w and Im w denote the real and imaginary parts of w. Equating the modulus of both sides of the equation $e^w = z$, we get

$$e^{\mathrm{Re}\ w} = |z|,$$

and thus

$$\mathrm{Re}\ w = \log(|z|).$$

Further, equating the argument of both sides, we get

$$\mathrm{Im}\ w = \theta = \arg z,$$

and thus

$$w = \log(|z|) + i\arg z.$$

We are thus led to define

$$\log(z) = \log(|z|) + i\arg z, \tag{33.10}$$

which extends the definition of the natural logarithm from the positive real numbers to non-zero complex numbers. We see that the imaginary part $\log(z)$ is not uniquely defined up to multiples of 2π, since $\arg z$ is not, and thus $\log(z)$ is *multi-valued*: the imaginary part of $\log(z)$ is not uniquely defined up to multiples of 2π. Choosing $\theta = \mathrm{Arg}\ z$ with $0 \leq \mathrm{Arg}\ z < 2\pi$, we obtain the *principal branch* of $\log(z)$ denoted by

$$\mathrm{Log}(z) = \log(|z|) + i\mathrm{Arg}\ z.$$

We see that if we let $\arg z$ increase from 0 beyond 2π, the function Log (z) will be discontinuous at Im $z = 2\pi$. We thus have to remember that the imaginary part of $\log(z)$ is not uniquely defined.

Chapter 33 Problems

33.1. Describe in geometrical terms the mappings $f : \mathbb{C} \to \mathbb{C}$ given by (a) $f(z) = \exp(z)$, (b) $f(z) = \mathrm{Log}(z)$, (c) $\sin(z)$.

34
Techniques of Integration

A poor head, having subsidiary advantages,.... can beat the best, just as a child can draw a line with a ruler better than the greatest master by hand. (Leibniz)

34.1 Introduction

It is not generally possible to find an explicit formula for a primitive function of a given arbitrary function in terms of known *elementary functions*, by which we mean the polynomials, rational functions, root functions, exponentials and trigonometric functions along with their inverses and combinations. It is not even true that the primitive function of an elementary function is another elementary function. A famous example is given by the function $f(x) = \exp(-x^2)$, whose primitive function $F(x)$ (with $F(0) = 0$), which exists by the Fundamental Theorem, is known *not* to be an elementary function (by a tricky proof by contradiction). To compute values of $F(x) = \int_0^x \exp(y)\, dy$ for different values of x we therefore have to use numerical quadrature just as in the case of the logarithm. Of course we can give $F(x)$ a *name*, for example we may agree to call it the *error function* $F(x) = \text{erf}(x)$ and add it to our list of known functions that we can use. Nevertheless there will be other functions (such as $\frac{\sin(x)}{x}$) whose primitive function cannot be expressed in the known functions.

The question of how to handle such functions (including also $\log(x)$, $\exp(x)$, $\sin(x)$...) of course arises: should we precompute long tables of values of these functions and print them in thick books or store them in the computer, or should we compute each required value from scratch using numerical quadrature? The first option was favoured in earlier times when computing power was sparse, and the second one is favoured today (even in the pocket calculator).

Despite the impossibility to reach generality, it is it is possible (and useful) to compute primitive functions analytically in certain cases, and in this chapter, we collect some tricks that have proved useful for doing this. The tricks we present are basically various clever substitutions together with integration by parts. We have no ambition to be encyclopedic. We refer to Mathematics Handbook for Science and Engineering for further development.

We start with rational functions, and then proceed to various combinations of polynomials, logarithms, exponentials and trigonometric functions.

34.2 Rational functions: the simple cases

Integration of rational functions depends on three basic formulas

$$\int_{x_0}^{x} \frac{1}{s-c}\,ds = \log|x-c| - \log|x_0-c|, \quad c \neq 0 \tag{34.1}$$

$$\int_{x_0}^{x} \frac{s-a}{(s-a)^2+b^2}\,dx = \frac{1}{2}\log((x-a)^2+b^2) - \frac{1}{2}\log((x_0-a)^2+b^2) \tag{34.2}$$

and

$$\int_{x_0}^{x} \frac{1}{(s-a)^2+b^2}\,ds = [\frac{1}{b}\arctan(\frac{x-a}{b})] - [\frac{1}{b}\arctan(\frac{x_0-a}{b})], \quad b \neq 0. \tag{34.3}$$

These formulas can be verified by differentiation. Using the formulas can be straightforward as in

Example 34.1.

$$\int_6^8 \frac{ds}{s-4} = \log 4 - \log 2 = \log 2.$$

Or more complicated as in

Example 34.2.

$$\begin{aligned}\int_2^4 \frac{ds}{2(s-2)^2+6} &= \frac{1}{2}\int_2^4 \frac{ds}{(s-2)^2+3}\\ &= \frac{1}{2}\int_2^4 \frac{ds}{(s-2)^2+(\sqrt{3})^2}\\ &= \frac{1}{2}\left(\frac{1}{\sqrt{3}}\arctan\left(\frac{4-2}{\sqrt{3}}\right) - \frac{1}{\sqrt{3}}\arctan\left(\frac{2-2}{\sqrt{3}}\right)\right).\end{aligned}$$

Of course we may combine these formulas with substitution:

EXAMPLE 34.3.

$$\int_0^x \frac{\cos(s)\,ds}{\sin(s)+2} = \int_0^{\sin(x)} \frac{du}{u+2} = \log|\sin(x)+2| - \log 2.$$

Using (34.2) and (34.3) may require *completing the square*, as we now show in

EXAMPLE 34.4. For example, consider

$$\int_0^3 \frac{ds}{s^2-2s+5}.$$

We want to get s^2-2s+5 into the form $(s-a)^2+b^2$ if possible. We set

$$(s-a)^2+b^2 = s^2-2as+a^2+b^2 = s^2-2s+5.$$

Equating the coefficients of s on both sides gives $a=1$. Equating the constant terms on both sides gives $b^2 = 5-1 = 4$ and therefore we may take $b=2$. After a little practice with completing the square, we can often argue directly, as

$$s^2-2s+5 = s^2-2s+1^2-1^2+5 = (s-1)^2+2^2.$$

Returning to the integral, we have

$$\int_0^3 \frac{ds}{s^2-2s+5} = \int_0^3 \frac{ds}{(s-1)^2+2^2} = \frac{1}{2}\arctan\left(\frac{3-2}{2}\right) - \frac{1}{2}\arctan\left(\frac{0-2}{2}\right).$$

34.3 Rational functions: partial fractions

We now investigate a systematic method for computing integrals of *rational* functions $f(x)$, i.e. functions of the form $f(x) = p(x)/q(x)$, where $p(x)$ and $q(x)$ are polynomials. The method is based manipulating the integrand so that the basic formulas (34.1)–(34.3) can be used. The manipulation is based on the observation that it is possible to write a complicated rational function as a sum of relatively simple rational functions.

EXAMPLE 34.5. Consider the integral

$$\int_4^5 \frac{s^2+s-2}{s^3-3s^2+s-3}\,ds.$$

The integrand can be expanded

$$\frac{s^2+s-2}{s^3-3s^2+s-3} = \frac{1}{s^2+1} + \frac{1}{s-3}$$

which we can verify by adding the two fractions on the right after computing a common denominator,

$$\begin{aligned}\frac{1}{s^2+1} + \frac{1}{s-3} &= \frac{s-3}{s-3} \times \frac{1}{s^2+1} + \frac{s^2+1}{s^2+1} \times \frac{1}{s-3} \\ &= \frac{s-3+s^2+1}{(s^2+1)(s-3)} = \frac{s^2+s-2}{s^3-3s^2+s-3}.\end{aligned}$$

Therefore we can integrate

$$\begin{aligned}\int_4^5 \frac{s^2+s-2}{s^3-3s^2+s-3}\, ds &= \int_4^5 \frac{1}{s^2+1}\, ds + \int_4^5 \frac{1}{s-3}\, ds \\ &= (\arctan(5) - \arctan(4)) + (\log(5-3) - \log(4-3)).\end{aligned}$$

The general technique of *partial fractions* is based on a systematic method for writing a rational function as a sum of simple rational functions that can be integrated with the basic formulas (34.1)–(34.3). The method is analogous to "reversing" the addition of rational functions by finding a common denominator.

Applying the technique of partial fractions to a general rational function has several steps, which we explain in "reverse" order. So we begin by assuming that the numerator $p(x)$ of the rational function $p(x)/q(x)$ has smaller degree than the denominator $q(x)$, i.e. $\deg p(x) < \deg q(x)$, and that $q(x)$ has the form

$$\frac{p(x)}{q(x)} = \frac{p(x)}{k(x-c_1)\cdots(x-c_n)((x-a_1)^2+b_1^2)\cdots((x-a_m)^2+b_m^2)}, \tag{34.4}$$

where k is a number, the c_i are the real roots of $q(x)$, and the second degree factors $(x-a_j)^2+b_j^2$ correspond to the complex roots $a_j \pm ib_j$ of $q(x)$ that necessarily come in pairs of complex conjugates. We call polynomials of the form $(x-a_j)^2+b_j^2$ *irreducible* because we cannot factor them as a product of linear polynomials with *real* coefficients.

In the first instance, we assume that the zeroes $\{c_i\}$ and $\{a_j \pm ib_j\}$ are distinct. In this case, we rewrite $p(x)/q(x)$ as the sum of partial fractions

$$\frac{p(x)}{q(x)} = \frac{C_1}{x-c_1} + \cdots + \frac{C_n}{x-c_n} + \frac{A_1(x-a_1)+B_1}{(x-a_1)^2+b_1^2} + \cdots + \frac{A_m(x-a_m)+B_m}{(x-a_m)^2+b_m^2}, \tag{34.5}$$

for some constants C_i, $1 \le j \le n$, and A_j, B_j, $1 \le j \le m$ that we have to determine. The motivation to rewrite $p(x)/q(x)$ in this way is that we can then compute an integral of $p(x)/q(x)$ by applying the formulas (34.1)–(34.3) to integrate the individual terms on the right-hand side of (34.5) as in the example above.

EXAMPLE 34.6. For $p(x) = q(x) = (x-1)/(x^2-x-2)$ with $q(x) = (x-2)(x+1)$ we have

$$\frac{x-1}{x^2-x-2} = \frac{x-1}{(x-2)(x+1)} = \frac{1/3}{x-2} + \frac{2/3}{x+1},$$

and thus

$$\begin{aligned}\int_{x_0}^{x} \frac{s-1}{s^2-s-2}\,ds &= \frac{1}{3}\int_{x_0}^{x}\frac{1}{s-2}\,ds + \frac{2}{3}\int_{x_0}^{x}\frac{1}{s+1}\,ds \\ &= \frac{1}{3}[\log(s-2)]_{s=x_0}^{s=x} + \frac{2}{3}[\log(s+1)]_{s=x_0}^{s=x}.\end{aligned}$$

The rationale for the expansion (34.5) is simply that if we ask for the most general sum of rational functions with denominators of degrees 1 and 2 that can yield $p(x)/q(x)$, where $q(x)$ is the common denominator for the sum, then we get precisely the right- hand side of (34.5). In particular if the terms on the right had numerators of any higher degree, then $p(x)$ would have to have degree greater than $q(x)$.

The constants C_i, A_j and B_j in (34.5) can be found by rewriting the right-hand side of (34.5) with a common denominator.

EXAMPLE 34.7. In the last example with $q(x) = (x-2)(x+1)$, we find that

$$\frac{C_1}{x-2} + \frac{C_2}{x+1} = \frac{C_1(x+1) + C_2(x-2)}{(x-2)(x+1)} = \frac{(C_1+C_2)x + (C_1-2C_2)}{(x-2)(x+1)},$$

which equals

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

if and only if

$$C_1 + C_2 = 1 \quad \text{and} \quad C_1 - 2C_2 = -1,$$

that is if $C_1 = 1/3$ and $C_2 = 2/3$.

Since it is cumbersome to compute the constants by dealing with the fractions, we usually rewrite the problem by multiplying both sides of (34.5) by the common denominator.

EXAMPLE 34.8. We multiply both sides of

$$\frac{x-1}{(x-2)(x+1)} = \frac{C_1}{x-2} + \frac{C_2}{x+1}$$

by $(x-2)(x+1)$ to get

$$x - 1 = C_1(x+1)C_2(x-2) = (C_1 + C_2)x + (C_1 - 2C_2).$$

Equating coefficients, we find $C_1 + C_2 = 1$ and $C_1 - 2C_2 = -1$, which yields $C_1 = 1/3$ and $C_2 = 2/3$.

EXAMPLE 34.9. To integrate $f(x) = (5x^2-3x+6)/((x-2)((x+1)^2+2^2))$, we begin by writing the partial fraction expansion

$$\frac{5x^2-3x+6}{(x-2)((x+1)^2+2^2)} = \frac{C}{x-2} + \frac{A(x+1)+B}{(x+1)^2+2^2}.$$

To determine the constants, we multiply both sides by $(x-2)((x+1)^2+2^2)$ to obtain

$$\begin{aligned}5x^2-3x+6 &= C((x+1)^2+2^2) + (A(x+1)+B)(x-2)\\ &= (C+A)x^2 + (2C-2A+B)x + (4C-2A-2B).\end{aligned}$$

Equating coefficients, we find that $C + A = 0$, $2C - 2A + B = 1$ and $5C - 2A - 2B = 0$, that is $C = 2$, $A = 3$ and $B = -1$. Therefore we find that

$$\begin{aligned}&\int_{x_0}^{x} \frac{5s^2-3s+6}{(s-2)((s+1)^2+2^2)}\,ds\\ &= 2\int_{x_0}^{x} \frac{1}{s-2}\,ds + \int_{x_0}^{x} \frac{3(s+1)-1}{(s+1)^2+2^2}\,ds\\ &= 2\int_{x_0}^{x} \frac{1}{s-2}\,ds + 3\int_{x_0}^{x} \frac{s+1}{(s+1)^2+2^2}\,ds - \int_{x_0}^{x} \frac{1}{(s+1)^2+2^2}\,ds\\ &= 2\big(\log|x-2| - \log|x_0-2|\big)\\ &\qquad + \frac{3}{2}\big(\log((x+1)^2+4) - \log((x_0+1)^2+4)\big)\\ &\qquad\quad - \frac{1}{2}\big(\arctan(\frac{x+1}{2}) - \arctan(\frac{x_0+1}{2})\big).\end{aligned}$$

In the case that some of the factors in the factorization of the denominator (34.4) are repeated, i.e. some of the roots have multiplicity greater than one, then we have to modify the partial sum expansion (34.5). We do not write out a general case because it is a mess and nearly unreadable, we just note that the principle for determining the correct partial fractions is always to write down the most general sum that can give the indicated common denominator.

EXAMPLE 34.10. The general partial fraction expansion of $f(x) = x^2/((x-2)(x+1)^2)$ has the form

$$\frac{x^2}{(x-2)(x+1)^2} = \frac{C_1}{x-2} + \frac{C_{2,1}}{x+1} + \frac{C_{2,2}}{(x+1)^2},$$

for constants C_1, $C_{2,1}$ and $C_{2,2}$ because all of the terms on the right-hand will yield the common denominator $(x-2)(x+1)^2$. Multiplying both sides by the common denominator and equating coefficients as usual, we find that $C_1 = 4/9$, $C_{2,1} = 5/9$ and $C_{2,2} = -3/9$.

In general if $q(x)$ has the multiple factor $(x-c_i)^L$ the term $\frac{C_i}{x-c_i}$ in the partial fraction expansion (34.5) should be replaced by the *sum* of fractions $\sum_{l=1}^{l=L} \frac{C_{i,l}}{(x-c)^l}$. There is a corresponding procedure for multiple factors of the form $((x-a)^2+b^2)^L$.

We have discussed how to integrate rational functions $p(x)/q(x)$ where $\deg p < \deg q$ and q is factored into a product of linear and irreducible quadratic polynomials. Now we discuss removing these restrictions. First we deal with the factorization of the denominator $q(x)$. The Fundamental Theorem of Algebra says that a polynomial q of degree n with real coefficients has exactly n roots and hence it can be factored into a product of n linear polynomials with possibly complex coefficients. However, because the polynomial q has real coefficients, the complex roots always come in complex conjugate pairs, i.e. if r is a root of q then so is $\bar{r}$. This means that there are an even number of linear factors of q corresponding to complex roots and furthermore we can combine the factors corresponding to conjugate roots to get quadratic factors with real coefficients. For example, $(x-3+i)(x-3-i) = (x-3)^2+1$. Therefore every polynomial $q(x)$ can theoretically be factored into a product $k(x-c_1)\cdots(x-c_n)((x-a_1)^2+b_1^2)\cdots((x-a_m)^2+b_m^2)$.

However, we caution that this theoretical result does not carry over in practice to situations in which the degree of q is large. To determine the factorization of q, we must determine the roots of q. In the problems and examples, we stick to cases in which the roots are simple, relatively small integers. But in general we know that the roots can be any kind of algebraic number which we can only approximate. Unfortunately it turns out that it is extremely difficult to determine the roots of a polynomial of high degree, even using Newton's method. So the method of partial fractions is used only for low degree polynomials in practice, though it is a very useful theoretical tool.

Finally we remove the restriction that $\deg p < \deg q$. When the degree of the numerator polynomial $p(x)$ is $\geq$ the degree of the denominator polynomial $q(x)$, we first use polynomial division to rewrite $f(x)$ as the sum of a polynomial $s(x)$ and a rational function $\frac{r(x)}{q(x)}$ for which the degree of the numerator $r(x)$ is *less* than the degree of the denominator $q(x)$.

EXAMPLE 34.11. For $f(x) = (x^3-x)/(x^2+x+1)$, we divide to get $f(x) = x-1+(1-x)/(x^2+x+1)$, so that

$$\int_0^{\bar{x}} \frac{x^3}{x^2+x+1}\,dx = [\frac{1}{2}x^2 - x]_{x=0}^{x=\bar{x}} + \int_0^{\bar{x}} \frac{1-x}{x^2+x+1}\,dx.$$

34.4 A product of a polynomial function and a trigonometric or exponential function

To integrate the product of a polynomial and a trigonometric or exponential function, we use integration by parts repeatedly to reduce the polynomial to an constant.

EXAMPLE 34.12. To compute a primitive function of $x\cos(x)$, we integrate by parts once

$$\int_0^x y\cos(y)\,dy = [y\sin(y)]_{y=0}^{y=x} - \int_0^x \sin(y)\,dy = x\sin(x) + \cos(x) + 1.$$

To handle higher order polynomials, we use integration by parts several times.

EXAMPLE 34.13. We have

$$\begin{aligned}\int_0^x s^2 e^s\,ds &= s^2(e^s)_{s=0}^{s=x} - 2\int_0^x se^s\,ds = [s^2e^s]_{s=0}^{s=x} - 2\big([se^s]_{s=0}^{s=x} - \int_0^x e^s\,ds\big)\\ &= [s^2e^s)_{s=0}^{s=x} - 2\big([se^s]_{s=0}^{s=x} - [e^s]_{s=0}^{s=x}\big) = x^2e^x - 2xe^x + 2e^x - 2.\end{aligned}$$

34.5 Combination of trigonometric and root functions

To compute a primitive function of $\sin(\sqrt{y})$ for $x > 0$, we set $y = t^2$ and obtain by using partial integration

$$\begin{aligned}\int_0^x \sin(\sqrt{y})\,dy &= \int_0^{\sqrt{x}} 2t\sin(t)\,dt = [-2t\cos(t)]_{t=0}^{t=\sqrt{x}} + 2\int_0^{\sqrt{x}} \cos(t)\,dt\\ &= -2\sqrt{x}\cos(\sqrt{x}) + 2\sin(\sqrt{x}).\end{aligned}$$

34.6 Products of exponential and trigonometric functions

To compute a primitive function of $e^y\sin(y)$, we use repeated integration by parts as follows

$$\begin{aligned}\int_0^x e^y\sin(y)\,dy &= [e^y\sin(y)]_{y=0}^{y=x} - \int_0^x e^y\cos(y)\,dy\\ &= e^x\sin(x) - [e^y\cos(y)]_{y=0}^{y=x} - \int_0^x e^y\sin(y)\,dy,\end{aligned}$$

which shows that

$$\int_0^x e^y \sin(y)\,dy = \frac{1}{2}(e^x \sin(x) - e^x \cos(x) + 1)$$

34.7 Products of polynomials and logarithms

To compute a primitive function of $x^2\log(x)$, we integrate by parts:

$$\int_1^x y^2\log(y)\,dy = [\frac{y^3}{3}\log(y)]_{y=1}^{y=x} - \int_1^x \frac{y^3}{3}\frac{1}{y}\,dy = \frac{x^3}{3}\log(x) - \frac{x^3}{9} + \frac{1}{9}.$$

Chapter 34 Problems

34.1. Compute (a) $\int_0^x t\sin(2t)\,dt$ (b) $\int_0^x t^2\cos(t)\,dt$ (c) $int_0^x t\exp(-2t)\,dt$. Hint: Integrate by parts.

34.2. Compute (a) $\int_1^x y\log(y)\,dy$ (b) $\int_1^x \log(y)\,dy$ (c) $\int_0^x \arctan(t)\,dt$ (d) $\int_0^x$ $exp(-t)\cos(2t)\,dt$. Hint: Integrate by parts.

34.3. Compute using the formula $\int_0^x \frac{g'(y)}{g(y)}\,dy = \log(g(x)) - \log(g(0))$ the following integrals. (a) $\int_0^x \frac{y}{y^2+1}\,dy$ (b) $\int_0^x \frac{e^t}{e^t+1}\,dt$.

34.4. Compute by a suitable change of variable (a) $\int_0^x y\exp(y^2)\,dy$ (b) $\int_0^x y\sqrt{y-1}\,dy$ (c) $\int_0^x \sin(t)\cos^2(t)\,dt$.

34.5. Compute (a) $\int_0^x \frac{dy}{y^2-y-2}\,dy$ (b) $\int_0^x \frac{y^3}{y^2+2y-3}\,dy$ (c) $\int_0^x \frac{dy}{y^2+2y+5}\,dy$ (d) $\int_0^x \frac{x-x^2}{(y-1)(y^2+2y+5)}\,dy$ (e) $\int_0^x \frac{x^4}{(x-1)(x^2+x-6)}\,dy$.

34.6. Recalling that a function is called *even* if $f(-x) = f(x)$ and *odd* if $f(-x) = -f(x)$ for all x, (a) give examples of even and odd functions (b) sketch their graphs, and (c) show that

$$\int_{-a}^a f(x)\,dx = 2\int_0^a f(x)\,dx \text{ if f is even,} \quad \int_{-a}^a f(x)\,dx = 0 \text{ if f is odd.} \quad (34.6)$$

34.7. Compute (a) $\int_{-\pi}^{\pi} |x|\cos(x)\,dx$ (b) $\int_{-\pi}^{\pi} \sin^2(x)\,dx$ (c) $\int_{-\pi}^{\pi} x\sin^2(x)\,dx$ (d) $\int_{-\pi}^{\pi} \arctan(x+3x^3)\,dx$.

35

Solving differential equations using the exponential

...he climbed a little further... and further...and then just a little further. (Winnie-the-Pooh)

35.1 Introduction

The exponential function plays a fundamental role in modeling and analysis because of its basic properties. In particular it can be used to solve a variety of differential equations analytically as we show in this chapter. We start with generalizations of the initial value problem (31.2) from Chapter *The exponential function*:

$$u'(x) = \lambda u(x) \quad \text{for } x > a, \quad u(a) = u_a, \tag{35.1}$$

where $\lambda \in \mathbb{R}$ is a constant, with solution

$$u(x) = \exp(\lambda(x-a))u_a \quad \text{for } x \geq a. \tag{35.2}$$

Analytic solutions formulas may give very important information and help the intuitive understanding of different aspects of a mathematical model, and should therfore be kept as valuable gems in the scientist/engineer's tool-bag. However, useful analytical formulas are relatively sparse and must be complemented by numerical solutions techniques. In Chapter *The General initial value problem* we extend the constructive numerical method for solving (35.1) to construct solutions of general initial value problems for systems of differential equations, capable of modeling a very large variety of phenomena. We can thus numerically compute the solution to just about any initial value problem, with more or less computational work, but we are limited to compting one solution for each specific choice of data, and getting qualitative information for a variety of different data may be costly. On the other hand, an analytical solution formula, when available, may contain this qualitative information for direct information.

An analytical solution formula for a differential equation may thus be viewed as a (smart and beautiful) short-cut to the solution, like evaluating an integral of a function by just evaluating two values of a corresponding primitive function. On the other hand, numerical solution of a differential equation is like a walk along a winding mountain road from point A to point B, without any short-cuts, similar to computing an integral by numerical quadrature. It is useful to be able to use both approaches.

35.2 Generalization to $u'(x) = \lambda(x)u(x) + f(x)$

The first problem we consider is a model in which the rate of change of a quantity $u(x)$ is proportional to the quantity with a variable factor of proportionality $\lambda(x)$, and moreover in which there is an external "forcing" function $f(x)$. The problem reads:

$$u'(x) = \lambda(x)u(x) + f(x) \quad \text{for } x > a, \quad u(a) = u_a, \tag{35.3}$$

where $\lambda(x)$ and $f(x)$ are given functions of x, and u_a is a given initial value. We first describe a couple physical situations being modeled by (35.3).

EXAMPLE 35.1. Consider for time $t > 0$ the population $u(t)$ of rabbits in West Virginia with inital value $u(0) = u_0$ given, which we assume has time dependent known birth rate $\beta(t)$ and death rate $\delta(t)$. In general, we would expect that rabbits will migrate quite freely back and forth across the state border and that the rates of the migration would vary with the season, i.e. with time t. We let $f_i(t)$ and $f_o(t)$ denote the rate of migration into and out of the state respectively at time t, which we assume to be known (realistic?). Then the population $u(t)$ will satisfy

$$\dot{u}(t) = \lambda(t)u(t) + f(t), \quad \text{for } t > a, \quad u(a) = u_a, \tag{35.4}$$

with $\lambda(t) = \beta(t) - \delta(t)$ and $f(t) = f_i(t) - f_o(t)$, which is of the form (35.3). Recall that $\dot{u} = \frac{du}{dt}$.

EXAMPLE 35.2. We model the amount of solute such as salt in a solvent such as water in a tank in which there is both inflow and outflow, see Fig. 35.1. We let $u(t)$ denote the amount of solute in the tank at time t and suppose that we know the initial amount u_0 at $t = 0$. We suppose that a mixture of solute/solvent, of concentration C_i in say grams per

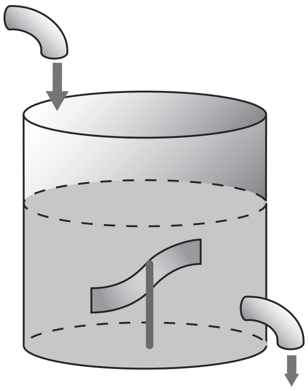


FIGURE 35.1. An illustration of a chemical mixing tank.

liter, flows into the tank at a rate σ_i liters per second. We assume there is also outflow at a rate of σ_o liters per second, and we assume that the mixture in the tank is well mixed with a uniform concentration $C(t)$ at any time t.

To get a differential equation for $u(t)$, we compute the change $u(t+\Delta t)-u(t)$ during the interval $[t, t+\Delta t]$. The amount of solute that flows into the tank during that time interval is $\sigma_i C_i \Delta t$, while the amount of solute that flows out of the tank during that time equals $\sigma_o C(t)\Delta t$, and thus

$$u(t+\Delta t) - u(t) \approx \sigma_i C_i \Delta t - \sigma_o C(t)\Delta t, \tag{35.5}$$

where the approximation improves when we decrease Δt. Now the concentration at time t will be $C(t) = u(t)/V(t)$ where $V(t)$ is the volume of fluid in the tank at time t. Substituting this into (35.5) and dividing by Δt gives

$$\frac{u(t+\Delta t) - u(t)}{\Delta t} \approx \sigma_i C_i - \sigma_o \frac{u(t)}{V(t)}$$

and taking the limit $\Delta t \to 0$ assuming $u(t)$ is differentiable gives the following differential equation for u,

$$\dot{u}(t) = -\frac{\sigma_o}{V(t)} u(t) + \sigma_i C_i.$$

The volume $V(t)$ is determined simply by the flow rates of fluid in and out of the tank. If there is initially V_0 liters in the tank then at time t, $V(t) = V_0 + (\sigma_i - \sigma_o)t$ because the flow rates are assumed to be constant. This gives again a model of the form (35.3):

$$\dot{u}(t) = -\frac{\sigma_o}{V_0 + (\sigma_i - \sigma_o)t} u(t) + \sigma_i C_i \quad \text{for } t > 0, \quad u(0) = u_0. \tag{35.6}$$

The method of integrating factor

We now return to derive an analytical solution formula for (35.3), using the method of *integrating factor*. To work out the solution formula, we begin with the special case

$$u'(x) = \lambda(x)u(x) \quad \text{for } x > a, \quad u(a) = u_a, \tag{35.7}$$

where $\lambda(x)$ is a given function of x. We let $\Lambda(x)$ be a primitive function of $\lambda(x)$ such that $\Lambda(a) = 0$, assuming that $\lambda(x)$ is Lipschitz continuous on $[a, \infty)$. We now multiply the equation $0 = u'(x) - \lambda(x)u(x)$ by $\exp(-\Lambda(x))$, we get

$$0 = u'(x)\exp(-\Lambda(x)) - u(x)\exp(-(x))\lambda(x) = \frac{d}{dx}(u(x)\exp(-\Lambda(x))),$$

where we refer to $\exp(-\Lambda(x))$ as an integrating factor because it brought the given equation to the form $\frac{d}{dx}$ of something, namely $u(x)\exp(-\Lambda(x))$, equal to zero. We conclude that $u(x)\exp(-\Lambda(x))$ is constant and is therefore equal to u_a since $u(a)\exp(-\Lambda(a)) = u(a) = u_a$. In other words, the solution to (35.7) is given by the formula

$$u(x) = \exp(\Lambda(x))u_a = e^{\Lambda(x)}u_a \quad \text{for } x \geq a. \tag{35.8}$$

We can check by differentiation that this function satisfies (35.7), and thus by uniqueness is the solution. To sum up, we have derived a solution formula for (35.7) in terms of the exponential function and a primitive function $\Lambda(x)$ of the coefficient $\lambda(x)$.

EXAMPLE 35.3. If $\lambda(x) = \frac{r}{x}$ and $a = 1$ then $\Lambda(x) = r\log(x) = \log(x^r)$, and the solution of

$$u'(x) = \frac{r}{x}u(x) \quad \text{for } x \neq 1, \quad u(1) = 1, \tag{35.9}$$

is according to (35.8) given by $u(x) = \exp(r\log(x)) = x^r$. We may define x^r for

Duhamel's principle

We now continue with the general problem to (35.3). We multiply by $e^{-\Lambda(x)}$, where again $\Lambda(x)$ is the primitive function of $\lambda(x)$ satisfying $\Lambda(a) = 0$, and get

$$\frac{d}{dx}(u(x)e^{-\Lambda(x)}) = f(x)e^{-\Lambda(x)}.$$

Integrating both sides, we see that the solution $u(x)$ satisfying $u(a) = u_a$ can be expressed as

$$u(x) = e^{\Lambda(x)}u_a + e^{\Lambda(x)}\int_a^x e^{-\Lambda(y)}f(y)\,dy. \tag{35.10}$$

This formula for the solution $u(x)$ of (35.3), expressing $u(x)$ in terms of the given data u_a and the primitive function $\Lambda(x)$ of $\lambda(x)$ satisfying $\Lambda(a) = 0$, is referred to as *Duhamel's principle* or the *variation of constants formula.*

We can check the validity of (35.10) by directly computing the derivative of $u(x)$:

$$\begin{aligned} u'(x) &= \lambda e^{\Lambda(x)} u_a + f(x) + \int_0^x (\lambda(x) e^{\Lambda(x)-\Lambda(y)} f(y)\, dy \\ &= \lambda(x) \left(e^{\Lambda(x)} u_a + \int_0^x e^{\Lambda(x)-\Lambda(y)} f(y)\, dy \right) + f(x). \end{aligned}$$

EXAMPLE 35.4. If $\lambda(x) = \lambda$ is constant, $f(x) = x$, $a = 0$ and $u_0 = 0$, the solution of (35.3) is given by

$$\begin{aligned} u(x) &= \int_0^x e^{\lambda(x-y)} y\, dy = e^{\lambda x} \int_0^x y e^{-\lambda y}\, dy \\ &= e^{\lambda x} \big([-\frac{y}{\lambda} e^{-\lambda y}]_{y=0}^{y=x} + \int_0^x \frac{1}{\lambda} e^{-\lambda y}\, dy \big) = -\frac{x}{\lambda} + \frac{1}{\lambda^2}(e^{\lambda x} - 1). \end{aligned}$$

EXAMPLE 35.5. In the model of the rabbit population (35.4), consider a situation with an initial population of 100, the death rate is greater than the birth rate by a constant factor 4, so $\lambda(t) = \beta(t) - \delta(t) = -4$, and there is a increasing migration into the state, so $f(t) = f_i(t) - f_o(t) = t$. Then (35.10) gives

$$\begin{aligned} u(t) &= e^{-4t} 100 + e^{-4t} \int_0^t e^{4s} s\, ds \\ &= e^{-4t} 100 + e^{-4t} \left(\frac{1}{4} s e^{4s} |_0^t - \frac{1}{4} \int_0^t e^{4s}\, ds \right) \\ &= e^{-4t} 100 + e^{-4t} \left(\frac{1}{4} t e^{4t} - \frac{1}{16} e^{4t} + \frac{1}{16} \right) \\ &= 100.0625 e^{-4t} + \frac{t}{4} - \frac{1}{16}. \end{aligned}$$

Without the migration into the state, the population would decrease exponentially, but in this situation the population decreases only for a short time before beginning to increase at a linear rate.

EXAMPLE 35.6. Consider a mixing tank in which the input flow at a rate of $\sigma_i = 3$ liters/sec has a concentration of $C_i = 1$ grams/liter, and the outflow is at a rate of $\sigma_o = 2$ liters/sec, the initial volume is $V_0 = 100$ liters with no solute desolved, so $u_0 = 0$. The equation is

$$\dot{u}(t) = -\frac{2}{100 + t} u(t) + 3.$$

We find $\Lambda(t) = 2\ln(100+t)$ and so

$$\begin{aligned}
u(t) &= 0 + e^{2\ln(100+t)} \int_0^t e^{-2\ln(100+s)} 3\,ds \\
&= (100+t)^2 \int_0^t (100+s)^{-2} 3\,ds \\
&= (100+t)^2 \left(\frac{-3}{100+t} + \frac{3}{100} \right) \\
&= \frac{3}{100} t(100+t).
\end{aligned}$$

As expected from the conditions, the concentration increases steadily until the tank is full.

35.3 The equation $u''(x) - u(x) = 0$

Consider the second order initial value problem

$$u''(x) - u(x) = 0 \quad \text{for } x > 0,\, u(0) = u_0,\, u'(0) = u_1, \tag{35.11}$$

with two initial conditions. We can write the differential equation $u''(x) - u(x) = 0$ formally as

$$(D+1)(D-1)u = 0,$$

where $D = \frac{d}{dx}$, since $(D+1)(D-1)u = D^2u - Du + Du - u = D^2u - u$. Setting $w = (D-1)u$, we thus have $(D+1)w = 0$, which gives $w(x) = ae^{-x}$ with $a = u_1 - u_0$, since $w(0) = u'(0) - u(0)$. Thus, $(D-1)u = (u_1 - u_0)e^{-x}$, so that by Duhamel's principle

$$\begin{aligned}
u(x) &= e^x u_0 + \int_0^x e^{x-y}(u_1 - u_0)e^{-y}\,dy \\
&= \frac{1}{2}(u_0 + u_1)e^x + \frac{1}{2}(u_0 - u_1)e^{-x}.
\end{aligned}$$

We conclude that the solution $u(x)$ of $u''(x) - u(x) = 0$ is a linear combination of e^x and e^{-x} with coefficients determined by the initial conditions. The technique of "factoring" the differential equation $(D^2 - 1)u = 0$ into $(D+1)(D-1)u = 0$, is very powerful and we now proceed to follow up this idea.

35.4 The differential equation $\sum_{k=0}^{n} a_k D^k u(x) = 0$

In this section, we look for solutions of the *linear differential equation with constant coefficients*:

$$\sum_{k=0}^{n} a_k D^k u(x) = 0 \quad \text{for } x \in I, \tag{35.12}$$

where the coefficients a_k are given real numbers, and I is a given interval. Corresponding to the *differential operator* $\sum_{k=0}^{n} a_k D^k$, we define the polynomial $p(x) = \sum_{k=0}^{n} a_k x^k$ in x of degree n with the same coefficients a_k as the differential equation. This is called the *characteristic polynomial* of the differential equation. We can now express the differential operator formally as

$$p(D)u(x) = \sum_{k=0}^{n} a_k D^k u(x).$$

For example, if $p(x) = x^2 - 1$ then $p(D)u = D^2 u - u$.

The technique for finding solutions is based on the observation that the exponential function $\exp(\lambda x)$ has the following property:

$$p(D)\exp(\lambda x) = p(\lambda)\exp(\lambda x), \tag{35.13}$$

which follows from repeated use of the Chain rule. This translates the differential operator $p(D)$ acting on $\exp(\lambda x)$ into the simple operation of multiplication by $p(\lambda)$. Ingenious, right?

We now seek solutions of the differential equation $p(D)u(x) = 0$ on an interval I of the form $u(x) = \exp(\lambda x)$. This leads to the equation

$$p(D)\exp(\lambda x) = p(\lambda)\exp(\lambda x) = 0, \quad \text{for } x \in I,$$

that is, λ should be a root of the polynomial equation

$$p(\lambda) = 0. \tag{35.14}$$

This algebraic equation is called the *characteristic equation* of the differential equation $p(D)u = 0$. To find the solutions of a differential equation $p(D)u = 0$ on the interval I, we are thus led to search for the roots $\lambda_1, \ldots.\lambda_n$, of the algebraic equation $p(\lambda) = 0$ with corresponding solutions $\exp(\lambda_1 x), \ldots, \exp(\lambda_n x)$. Any linear combination

$$u(x) = \alpha_1 \exp(\lambda_1) + \ldots. + \alpha_n \exp(\lambda_n), \tag{35.15}$$

with α_i real (or complex) constants, will then be a solution of the differential equation $p(D)u = 0$ on I. If there are n distinct roots $\lambda_1, \ldots, \lambda_n$, then the *general solution* of $p(D)u = 0$ has this form. The constants α_i will be determined from initial or boundary conditions in a specific situation.

If the equation $p(\lambda) = 0$ has a multiple roots λ_i of multiplicity r_i, then the situation is more complicated. It can be shown that the solution is a sum of terms of the form $q(x)\exp(\lambda_i x)$, where $q(x)$ is a polynomial of degree at most $r_i - 1$. For example, if $p(D) = (D-1)^2$, then the general solution of $p(D)u = 0$ has the form $u(x) = (\alpha_0 + \alpha_1 x)\exp(x)$. In the Chapter *N-body systems* below we study the the constant coefficient linear second order equation $a_0 + a_1 Du + a_2 D^2 u = 0$ in detail, with interesting results!

The translation from a differential equation $p(D)u = 0$ to an algebraic equation $p(\lambda) = 0$ is very powerful, but requires the coefficients a_k of $p(D)$ to be independent of x and is thus not very general. The whole branch of *Fourier analysis* is based on the formula (35.13).

EXAMPLE 35.7. The characteristic equation for $p(D) = D^2 - 1$ is $\lambda^2 - 1 = 0$ with roots $\lambda_1 = 1, \lambda_2 = -1$, and the corresponding general solution is given by $\alpha_1 \exp(x) + \alpha_2 \exp(-x)$. We already met this example just above.

EXAMPLE 35.8. The characteristic equation for $p(D) = D^2 + 1$ is $\lambda^2 + 1 = 0$ with roots $\lambda_1 = i, \lambda_2 = -i$, and the corresponding general solution is given by

$$\alpha_1 \exp(ix) + \alpha_2 \exp(-ix).$$

with the α_i complex constants. Taking the real part, we get solutions of the form

$$\beta_1 \cos(x) + \beta_2 \sin(x)$$

with the β_i real constants.

35.5 The differential equation $\sum_{k=0}^{n} a_k D^k u(x) = f(x)$

Consider now the nonhomogeneous differential equation

$$p(D)u(x) = \sum_{k=0}^{n} a_k D^k u(x) = f(x), \tag{35.16}$$

with given constant coefficients a_k, and a given right hand side $f(x)$. Suppose $u_p(x)$ is any solution of this equation, which we refer to as a *particular solution*. Then any other solution $u(x)$ of $p(D)u(x) = f(x)$ can be written

$$u(x) = u_p(x) + v(x)$$

where $v(x)$ is a solution of the corresponding homogeneous differential equation $p(D)v = 0$. This follows from linearity and uniqueness since $p(D)(u - u_p) = f - f = 0$.

EXAMPLE 35.9. Consider the equation $(D^2 - 1)u = f(x)$ with $f(x) = x^2$. A particular solution is given by $u_p(x) = -x^2 - 2$, and thus the general solution is given by

$$u(x) = -x^2 - 2 + \alpha_1 \exp(x) + \alpha_2 \exp(-x).$$

FIGURE 35.2. Leonard Euler: "... I soon found an opportunity to be introduced to a famous professor Johann Bernoulli . ... True, he was very busy and so refused flatly to give me private lessons; but he gave me much more valuable advice to start reading more difficult mathematical books on my own and to study them as diligently as I could; if I came across some obstacle or difficulty, I was given permission to visit him freely every Sunday afternoon and he kindly explained to me everything I could not understand ... ".

35.6 Euler's differential equation

In this section, we consider Euler's equation

$$a_0 u(x) + a_1 x u'(x) + a_2 x^2 u''(x) = 0, \tag{35.17}$$

which has variable coeffients $a_i x^i$ of a very particular form. Following a grand mathematical tradition, we guess, or make an *Ansatz* on the form of the solution, and assume that $u(x) = x^m$ for some m to be determined. Substituting into the differential equation, we get

$$a_0 x^m + a_1 x (x^m)' + a_2 x^2 (x^m)'' = (a_0 + (a_1 - 1)m + a_2 m^2) x^m,$$

and we are thus led to the auxiliary algebraic equation

$$a_0 + (a_1 - 1)m + a_2m^2 = 0$$

in m. Letting the roots of this equation be m_1 and m_2, assuming the roots are real, any linear combination

$$\alpha_1 x^{m_1} + \alpha_2 x^{m_2}$$

is a solution of (35.17). In fact the general solution of (35.17) has this form if m_1 and m_2 are distint and real.

EXAMPLE 35.10. The auxiliary equation for the differential equation $x^2u'' - \frac{3}{2}xu' - 2u = 0$ is $m^2 - \frac{7}{2}m - 2 = 0$ with roots $m_1 = -\frac{1}{2}$ and $m_2 = 4$ and thus the general solution takes the form

$$u(x) = \alpha_1 \frac{1}{\sqrt{x}} + \alpha_2 x^4.$$

Leonard Euler (1707-83) is the mathematical genius of the 18th century, with an incredible production of more than 800 scientific articles half of them written after he got completely blind 1766, see Fig. 35.2.

Chapter 35 Problems

35.1. Solve the initial value problem (35.7) with $\lambda(x) = x^r$, where $r \in \mathbb{R}$, and $a = 0$.

35.2. Solve the following initial value problems: a) $u'(x) = 8xu(x), \quad u(0) = 1, \quad x > 0$, b) $\frac{(15x+1)u(x)}{u'(x)} = 3x, \quad u(1) = e, \quad x > 1$, c) $u'(x) + \frac{x}{(1-x)(1+x)}u = 0, \quad u(0) = 1, \quad x > 0$.

35.3. Make sure that you got the correct answer in the previous problem, part c). Will your solution hold for $x > 1$ as well as $x < 1$?

35.4. Solve the following initial value problems: a) $xu'(x) + u(x) = x, \quad u(1) = \frac{3}{2}, \quad x > 1$, b) $u'(x) + 2xu = x, \quad u(0) = 1, \quad x > 0$, c) $u'(x) = \frac{x+u}{2}, \quad u(0) = 0, \quad x > 0$.

35.5. Describe the behavior of the population of rabbits in West Virginia in which the birth rate exceeds the death rate by 5, the initial population is 10000 rabbits, and (a) there is a net migration out of the state at a rate of $5t$ (b) there is a net migration out of the state at a rate of $\exp(6t)$.

35.6. Describe the concentration in a mixing tank with an initial volume of 50 liters in which 20 grams of solute are dissolved, there is an inflow of 6 liters/sec with a concentration of 10 grams/liter and an outflow of 7 liters/sec.

36
Improper integrals

> All sorts of funny thoughts, run around my head. (When We Were Very Young, Milne)

36.1 Introduction

In some applications, it is necessary to compute integrals of functions that are unbounded at isolated points or to compute integrals of functions over unbounded intervals. We call such integrals *improper*, or sometimes (more properly) *generalized* integrals. We compute these integrals using the basic results on convergence of sequences that we have already developed.

We now consider these two kinds of improper integrals: integrals over unbounded intervals and integrals of unbounded functions.

36.2 Integrals over unbounded intervals

We start considering the following example of an integral over the unbounded interval $[0, \infty)$:

$$\int_0^\infty \frac{1}{1+x^2}\, dx.$$

The integrand $f(x) = (1+x^2)^{-1}$ is a smooth (positive) function that we can integrate over any finite interval $[0, n]$ to get,

$$\int_0^n \frac{1}{1+x^2}\,dx = \arctan(n). \tag{36.1}$$

Now we consider what happens as n increases, that is we integrate f over increasingly longer intervals. Since $\lim_{n\to\infty} \arctan(n) = \pi/2$, we may write

$$\lim_{n\to\infty} \int_0^n \frac{1}{1+x^2}\,dx = \frac{\pi}{2},$$

and we are thus led to *define*

$$\int_0^\infty \frac{1}{1+x^2}\,dx = \lim_{n\to\infty} \int_0^n \frac{1}{1+x^2}\,dx = \frac{\pi}{2}.$$

We generalize in the obvious way to an arbitrary (Lipschitz continuous) function $f(x)$ defined for $x > a$, and thus define

$$\int_a^\infty f(x)\,dx = \lim_{n\to\infty} \int_a^n f(x)\,dx \tag{36.2}$$

granted that the limit is defined and is finite. In this case, we say the improper integral is *convergent* (or is *defined*) and that the function $f(x)$ is *integrable* over $[a, \infty)$. Otherwise, we say the integral is *divergent* (or is *undefined*), and that $f(x)$ is *not* integrable over $[a, \infty)$.

If the function $f(x)$ is positive, then in order for the integral $\int_a^\infty f(x)\,dx$ to be convergent, the integrand $f(x)$ has to get sufficiently small for large values of x, since otherwise $\lim_{n\to\infty} \int_a^n f(x)\,dx = \infty$ and the integral is divergent. We saw above that the function $\frac{1}{1+x^2}$ was decaying to zero sufficiently quickly for large values of x to be integrable over $[a, \infty)$.

Consider now the function $\frac{1}{1+x}$ with a less quick decay as $x \to \infty$. Is it integrable on $[0, \infty)$? Well, we have

$$\int_0^n \frac{1}{1+x}\,dx = \Big[\log(1+x)\Big]_0^n = \log(1+n),$$

and since

$$\log(1+n) \to \infty \quad \text{as } n \to \infty$$

although the divergence is slow, we understand that $\int_0^\infty \frac{1}{1+x}\,dx$ is divergent.

EXAMPLE 36.1. The improper integral

$$\int_1^\infty \frac{dx}{x^\alpha}$$

is convergent for $\alpha > 1$, since

$$\lim_{n\to\infty} \int_1^n \frac{dx}{x^{-\alpha}} = \lim_{n\to\infty} [-\frac{x^{-(\alpha-1)}}{\alpha-1}]_1^n = \frac{1}{\alpha-1}.$$

We can sometimes show that an improper integral exists even when we can not compute its value.

EXAMPLE 36.2. Consider the improper integral

$$\int_1^\infty \frac{e^{-x}}{x}\,dx.$$

Since $f(x) = \frac{e^{-x}}{x} > 0$ for $x > 1$, we see that the sequence $\{I_n\}_{n=1}^\infty$, with

$$I_n = \int_1^n \frac{e^{-x}}{x}\,dx$$

is increasing. By Chapter *Optimization* we know that $\{I_n\}_{n=1}^\infty$ will have a limit if we only can show that $\{I_n\}_{n=1}^\infty$ is bounded above. Since trivially $1/x \le 1$ if $x \le 1$, we have for all $n \ge 1$

$$I_n \le \int_1^n e^{-x}\,dx = e^{-1} - e^{-n} \le e^{-1}.$$

We conclude that $\int_1^\infty \frac{e^{-x}}{x}\,dx$ converges. Note that we may restrict n to take integer values because the integrand e^{-x}/x tends to zero as x tends to infinity.

We may also compute integrals of the form

$$\int_{-\infty}^\infty f(x)\,dx.$$

We do this by choosing an arbitrary point $-\infty < a < \infty$ and defining

$$\begin{aligned}\int_{-\infty}^\infty f(x)\,dx &= \int_{-\infty}^a f(x)\,dx + \int_a^\infty f(x)\,dx \\ &= \lim_{m\to-\infty}\int_m^a f(x)\,dx + \lim_{n\to\infty}\int_a^n f(x)\,dx,\end{aligned} \tag{36.3}$$

where we compute the two limits independently and both must be defined and finite for the integral to exist.

36.3 Integrals of unbounded functions

We begin this section by considering the integral

$$\int_a^b f(x)\,dx,$$

where $f(x)$ is unbounded at a, i.e. $\lim_{x \downarrow a} f(x) = \pm\infty$. We consider the following example:

$$\int_0^1 \frac{1}{\sqrt{x}}\, dx.$$

The function $\frac{1}{\sqrt{x}}$ is unbounded on $(0,1]$, but bounded and Lipschitz continuous on $[\epsilon, 1]$ for any $1 \geq \epsilon > 0$. This means that the integrals

$$I_\epsilon = \int_\epsilon^1 \frac{1}{\sqrt{x}}\, dx = 2 - 2\sqrt{\epsilon} \tag{36.4}$$

are defined for any $1 \geq \epsilon > 0$, and evidently

$$\lim_{\epsilon \downarrow 0 \to \infty} I_\epsilon = 2,$$

where we recall that $\epsilon \downarrow 0$ means that ϵ tends to zero through positive values. It is thus natural to define

$$\int_0^1 \frac{1}{\sqrt{x}}\, dx = \lim_{\epsilon \downarrow 0} \int_\epsilon^1 \frac{1}{\sqrt{x}}\, dx = 2.$$

In general if $f(x)$ is unbounded close to a, then we define

$$\int_a^b f(x)\, dx = \lim_{s \downarrow a} \int_s^b f(x)\, dx, \tag{36.5}$$

and if $f(x)$ is unbounded at b then we define

$$\int_a^b f(x)\, dx = \lim_{s \uparrow a} \int_a^s f(x)\, dx \tag{36.6}$$

when these limits are defined and finite. As above, we say the improper integrals are convergent and defined if the limits exist and are finite, and otherwise say the integrals are divergent and not defined.

We may naturally extend this definition to the case when $f(x)$ is unbounded at a point $a < c < b$ by defining

$$\begin{aligned} \int_a^b f(x)\, dx &= \lim \int_a^c f(x)\, dx + \lim \int_c^b f(x)\, dx \\ &= \lim_{s \uparrow c} \int_a^s f(x)\, dx + \lim_{t \downarrow c} \int_t^b f(x)\, dx \end{aligned} \tag{36.7}$$

where the two limits are computed imdependently and must both be defined and finite for the integral to converge.

Chapter 36 Problems

36.1. If possible, compute the following integrals

1. $\displaystyle\int_0^\infty \frac{x}{(1+x^2)^2}\,dx$
2. $\displaystyle\int_{-\infty}^\infty xe^{-x^2}\,dx$
3. $\displaystyle\int_0^1 \frac{1}{\sqrt{1-x}}\,dx$
4. $\displaystyle\int_0^\pi \frac{\cos(x)}{(1-\sin(x))^{1/3}}\,dx$

36.2. Prove that if $\int_0^\infty |f(x)|\,dx$ is convergent, then so is $\int_0^\infty f(x)\,dx$, that is, absolute convergence implies convergence.

36.3. Prove that $\int_B \|x\|^{-\alpha}\,dx$, where $B = \{x \in \mathbb{R}^d : \|x\| < 1\}$, is convergent if $\alpha < d$ for $d = 1, 2, 3$.

37
Series

If you disregard the very simplest cases, there is in all of mathematics not a single series whose sum has been rigorously determined. In other words, the most important part of mathematics stand without a foundation.(Abel 1802-1829)

37.1 Introduction

In this chapter we consider the concept of *series*, which is a sum of numbers. We distinguish between a *finite* series, where the sum has a finite number of terms, and an *infinite series* with an infinite number of terms. A finite series does not pose any mysteries; we can, at least in principle, compute the sum of a finite series by adding the terms one-by-one, given enough time. The concept of an infinite series requires some explanation, since we cannot actually add an infinite number of terms one-by-one, and we thus need to define what we mean by an "infinite sum".

The concept of infinite series has a central role in Calculus, because a basic idea has been to seek to express "arbitrary" functions in terms of series as sums of simple terms. This was the grand idea of Fourier who thought of representing general functions as sums of trigonometric functions in the form of Fourier series, and Weierstrass who tried to do the same with monomials or polynomials in the form of power series. There are limitations to both Fourier and power series and the role of such series is today largely being taken over by computational methods. We therefore do not go into any excessive treatment of series, but we do present some important basic facts, which are useful to know.

We recall that we already met one infinite series, namely the *geometric series*

$$\sum_{i=0}^{\infty} a^i = 1 + a + a^2 + a^3 + \cdots,$$

where a is a real number. We determined the sum of this infinite series in the case $|a| < 1$ by first computing the *partial sum of order* n:

$$s_n = \sum_{i=0}^{n} a^i = 1 + a + a^2 + \cdots + a^n = \frac{1 - a^{n+1}}{1 - a}.$$

by summing the terms a^i with $i \leq n$. We then made the observation that if $|a| < 1$, then

$$\lim_{n\to\infty} s_n = \lim_{n\to\infty} \frac{1 - a^{n+1}}{1 - a} = \frac{1}{1 - a},$$

and so we defined for $|a| < 1$ the sum of the infinite geometric series to be

$$\sum_{i=0}^{\infty} a^i = \lim_{n\to\infty} \sum_{i=0}^{n} a^i = \frac{1}{1 - a}.$$

We note that if $|a| \geq 1$, then we had to leave the sum of the geometric series $\sum_{i=0}^{\infty} a^i$ undefined. If $|a| \geq 1$, then $|s_n - s_{n-1}| = |a^n| \geq 1$, and therefore $\{s_n\}_{n=0}^{\infty}$ is not a Cauchy sequence, and thus $\lim_{n\to\infty} s_n = \lim_{n\to\infty} \sum_{i=0}^{n} a^i$ does not exist. Evidently, a necessary condition for convergence is that the terms a^i tend to zero as i tends to infinity.

37.2 Definition of convergent infinite series

We now generalize these ideas to arbitrary infinite series. Thus let $\{a_n\}_{n=0}^{\infty}$ denote a sequence of real numbers and consider the *sequence of partial sums* $\{s_n\}_{n=0}^{\infty}$, where

$$s_n = \sum_{i=0}^{n} a_i = a_0 + a_1 + \cdots + a_n \tag{37.1}$$

is the *partial sum of order* n. We now say that the series $\sum_{i=0}^{\infty} a_i$ is *convergent* if the corresponding sequence of partial sums $\{s_n\}_{n=0}^{\infty}$ converges, and we then write

$$\sum_{i=0}^{\infty} a_i = \lim_{n\to\infty} s_n = \lim_{n\to\infty} \sum_{i=0}^{n} a_i, \tag{37.2}$$

which we refer to as the sum of the series. The convergence of a series $\sum_{i=1}^{\infty} a_i$ is thus reduced to the convergence of the sequence of it's partial sums. All convergence issues for a series are handled in this way by reduction to convergence of sequences. This chapter therfore may be viewed as a direct gives a direct continuation of Chapters *Sequences and limits* and *Real numbers*. In particular, we understand as in the case of a geometric series, that a necessary condition for convergence of a series $\sum_{i=0}^{\infty} a_i$ is that the terms a_i tend to zero as i tends to infinity. However, this condition is not sufficient, as we should know from our previous experience with sequences, and as we will see again below.

Note that we can similarly consider series of the form $\sum_{i=1}^{\infty} a_i$, or $\sum_{i=m}^{\infty} a_i$ for any integer m.

Note that in a few special cases like the geometric series, we can actually find an analytic formula for the sum of the series. However, for most series $\sum_{i=0}^{\infty} a_i$ this is not possible, or may be so be tricky that we can't make it. Of course, we can then usually compute an approximation by directly computing a partial sum $s_n = \sum_{i=0}^{n} a_i$ for some appropriate n, that is, if n is not too big and the terms a_i not too difficukt to evaluate. To then estimate the error, we are led to estimate the remainder $\sum_{i=n+1}^{\infty} a_i$. Thus we see a need to be able to analytically estimate the sum of a series, which may be easier than to analytically compute the exact sum.

In particular, such estimation may be used to decide if a series is convergent or not, which of course is an important issue because playing around with divergent series cannot have any meaning. In this pursuit, it is natural to distinguish between series in which all of the terms have the same sign and those in which the terms can have different signs. It may be more difficult to determine convergence for a series in which the terms can have different signs because of the possibility of cancellation between the terms.

Further, if we bound a series remainder $\sum_{i=n+1}^{\infty} a_i$ by using the triangle inequality, we get

$$|\sum_{i=n+1}^{\infty} a_i| \leq \sum_{i=n+1}^{\infty} |a_i|,$$

where the series on the right hand side is positive. So, positive series are of prime importance and we now turn to this topic.

37.3 Positive series

A series $\sum_{i=1}^{\infty} a_i$ is said to be a *positive series*, if $a_i \geq 0$ for $i = 1, 2, ...$ The important point about a positive series is that the sequence of partial sums is non-decreasing, because

$$s_{n+1} - s_n = \sum_{i=1}^{n+1} a_i - \sum_{i=1}^{n} a_i = a_{n+1} \geq 0. \tag{37.3}$$

In Chapter *Optimization* below we shall prove that a nondecreasing sequence converges if and only if the sequence is bounded above. If we accept this as a fact, we understand that a positive series is convergent if and only if the sequence of partial sums is bounded above, that is there is a constant C such that

$$\sum_{i=1}^{n} a_i \leq C \quad \text{for } n = 1, 2, ...,. \tag{37.4}$$

This gives a definite way to check convergence, which we state as a theorem:

Theorem 37.1 *A positive series converges if and only if the sequence of partial sums is bounded above.*

This result does not apply if the series has terms with different signs. For example, the series $\sum_{i=0}^{\infty}(-1)^i = 1 - 1 + 1 - 1 + 1...$ has bounded partial sums, but is not convergent since $(-1)^i$ does not tend to zero as i tends to infinity.

EXAMPLE 37.1. We can sometimes use an integral to bound the partial sums of a positive series and thus to prove convergence or estimate remainders. As an example, consider the positive series $\sum_{i=2}^{\infty} \frac{1}{i^2}$. The partial sum

$$s_n = \sum_{i=2}^{n} \frac{1}{i^2}$$

may be viewed as a quadrature formula for the integral of $\int_1^n x^{-2}\, dx$, see Fig. 37.1. MOre precisley, we see that

$$\begin{aligned}
\int_1^n x^{-2}\, dx &= \int_1^2 x^{-2}\, dx + \int_2^3 x^{-2}\, dx + \cdots + \int_{n-1}^n x^{-2}\, dx \\
&\geq \int_1^2 \frac{1}{2^2}\, dx + \int_2^3 \frac{1}{3^2}\, dx + \cdots + \int_{n-1}^n \frac{1}{n^2}\, dx \\
&\geq \frac{1}{2^2} + \frac{1}{3^2} + \cdots + \frac{1}{n^2} = s_n.
\end{aligned}$$

Since

$$\int_1^n x^{-2}\, dx = \left(1 - \frac{1}{n}\right) \leq 1,$$

we conclude that $s_n \leq 1$ for all n and therefore the series $\sum_{i=2}^{\infty} \frac{1}{i^2}$ is convergent. To compute an approximation of the sum of the series, we of course compute a partial sum s_n with n sufficiently large. To estimate the remainder we may of course use a similar comparison, see Problem 37.5.

EXAMPLE 37.2. The positive series $\sum_{i=1}^{\infty} \frac{1}{i+i^2}$ converges because for all n

$$s_n = \sum_{i=1}^{n} \frac{1}{i+i^2} \leq \sum_{i=1}^{n} \frac{1}{i^2} \leq 2$$

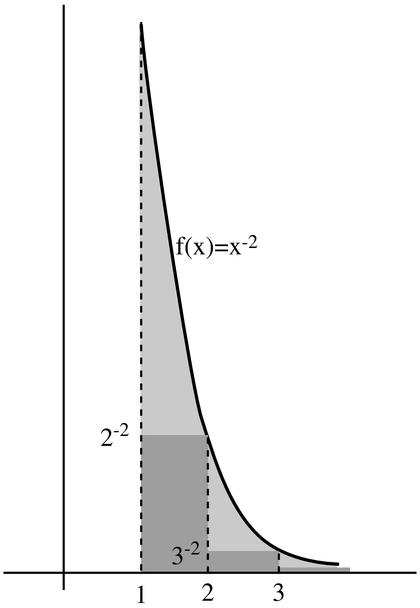


FIGURE 37.1. The relation between $\int_1^n x^{-2}\,dx$ and $\sum_{i=2}^n \frac{1}{i^2}$.

by the previous example.

Similarly, a *negative series* with all terms non-positive, converges if and only if its partial sums are bounded *below*.

EXAMPLE 37.3. For the *alternating* series

$$\sum_{i=1}^{\infty} \frac{(-1)^i}{i},$$

we have that the difference between two successive partial sums

$$s_n - s_{n-1} = \frac{(-1)^n}{n}$$

alternates in sign, and thus the sequence of partial sums is not monotone, and therefore we cannot decide convergence or not from the above theorem. We shall return to this series below and prove that it is in fact convergent.

37.4 Absolutely convergent series

Now we turn to series with terms of different signs. We begin by first considering series that converge regardless of any cancellation between the terms. We are motivated by the convergence results for positive series. A series $\sum_{i=1}^{\infty} a_i$ is said to be *absolutely convergent* if the series

$$\sum_{i=1}^{\infty} |a_i|$$

converges. By the previous result we know that a series $\sum_{i=1}^{\infty} a_i$ is absolutely convergent if and only if the sequence $\{\hat{s}_n\}$ with

$$\hat{s}_n = \sum_{i=1}^{n} |a_i|, \tag{37.5}$$

is bounded above.

We shall now prove that an absolutely convergent series $\sum_{i=1}^{\infty} a_i$ is convergent. By the triangle inequality we have for $m > n$,

$$|s_m - s_n| = |\sum_{n}^{m} a_i| \leq \sum_{n}^{m} |a_i| = |\hat{s}_m - \hat{s}_n|. \tag{37.6}$$

Now, since we can make $|\hat{s}_m - \hat{s}_n|$ arbitrarily small by taking m and n large, because $\sum_{i=1}^{\infty} a_i$ is absolutely convergent and thus $\{\hat{s}_n\}_{n=1}^{\infty}$ is a Cauchy sequence, we conclude that $\{s_n\}_{n=1}^{\infty}$ is a Cauchy sequence and therefore converges and thus the series $\sum_{i=1}^{\infty} a_i$ is convergent. We state this fundamental result as a theorem:

Theorem 37.2 *An absolutely convergent series is convergent.*

EXAMPLE 37.4. The series $\sum_{i=1}^{\infty} \frac{(-1)^i}{i^2}$ is convergent because $\sum_{i=1}^{\infty} \frac{1}{i^2}$ is convergent.

37.5 Alternating series

The convergence of a general series with terms of "random" sign may be very difficult to analyze because of cancellation of terms. We now consider a special case with a regular pattern to the signs of the terms:

$$\sum_{i=0}^{\infty} (-1)^i a_i \tag{37.7}$$

where $a_i \geq 0$ for all i. This is called an *alternating series* since the signs of the terms alternate. We shall now prove that if $a_{i+1} \leq a_i$ for $i = 0, 1, 2...$ and $\lim_{i\to\infty} a_i = 0$, then the alternating series converges. The key observation is that the sequence $\{s_n\}$ of partial sums satisfies

$$s_1 \leq s_3 \leq s_5 \leq ...s_{2j+1} \leq s_{2i} \leq ... \leq s_4 \leq s_2 \leq s_0, \tag{37.8}$$

which shows that both limits $\lim_{j\to\infty} s_{2j+1}$ and $\lim_{i\to\infty} s_{2i}$ exist. Since $a_i \to 0$ as i tends to infinity, $\lim_{j\to\infty} s_{2j+1} = \lim_{i\to\infty} s_{2i}$, and thus $\lim_{n\to\infty} s_n$ exists and convergence of the series $\sum_{i=0}^{\infty}(-1)^i a_i$ follows. We summarize in the following theorem first stated and proved by Leibniz:

Theorem 37.3 *An alternating series with the property that the modulus of its terms tends monotonically to zero, converges.*

Example 37.5. The *harmonic series*

$$\sum_{i=1}^{\infty} \frac{(-1)^{i-1}}{i} = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \cdots$$

converges. We now proceed to show that this series is not absolutely convergent.

37.6 The series $\sum_{i=1}^{\infty} \frac{1}{i}$ theoretically diverges!

We shall now show that the *harmonic series* $\sum_{i=1}^{\infty} \frac{(-1)^i}{i}$ is **not** absolutely convergent, i.e. we shall prove that the series

$$\sum_{i=1}^{\infty} \frac{1}{i} = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$$

diverges. We do this by proving that the sequence $\{s_n\}_{n=1}^{\infty}$ of partial sums

$$s_n = \sum_{i=1}^{n} \frac{1}{i}$$

can become arbitrarily large if n is large enough. To see this we group the terms of a partial sum as follows:

$$\begin{aligned}
1 + \overline{\frac{1}{2}} + \overline{\frac{1}{3} + \frac{1}{4}} + \overline{\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}}& \\
+ \overline{\frac{1}{9} + \frac{1}{10} + \frac{1}{11} + \frac{1}{12} + \frac{1}{13} + \frac{1}{14} + \frac{1}{15} + \frac{1}{16}}& \\
+ \overline{\frac{1}{17} + \cdots + \frac{1}{32}} + \cdots&
\end{aligned}$$

The first "group" is $1/2$. The second group is

$$\frac{1}{3} + \frac{1}{4} \geq \frac{1}{4} + \frac{1}{4} = \frac{1}{2}.$$

The third group is

$$\frac{1}{5}+\frac{1}{6}+\frac{1}{7}+\frac{1}{8}\geq\frac{1}{8}+\frac{1}{8}+\frac{1}{8}+\frac{1}{8}=\frac{1}{2}.$$

The fourth group

$$\frac{1}{9}+\frac{1}{10}+\frac{1}{11}+\frac{1}{12}+\frac{1}{13}+\frac{1}{14}+\frac{1}{15}+\frac{1}{16}$$

has 8 terms that are larger than $1/16$, so it also gives a sum larger than $8/16 = 1/2$. We can continue in this way, taking the next 16 terms, all of which are larger than $1/32$, then the next 32 terms, all of which are larger than $1/64$, and so on. Each time we take a group, we get a contribution to the overall sum that is larger than $1/2$.

When we take n larger and larger, we can combine more and more terms in this way, making the sum larger in increments of $1/2$ each time. The partial sums therefore just become larger and larger as n increases, which means the partial sums diverge to infinity.

Note that by the arithmetic rules, the partial sum s_n should be the same whether we compute the sum in the "forward" direction

$$s_n = 1+\frac{1}{2}+\frac{1}{3}+\cdots\frac{1}{n-1}+\frac{1}{n}$$

or the "backward" direction

$$s_n = \frac{1}{n}+\frac{1}{n-1}+\cdots+\frac{1}{3}+\frac{1}{2}+1.$$

In Fig. 37.2, we list various partial sums in both the forward and backward directions computed using FORTRAN with single precision variables with about 7 digits of accuracy. Note two things about these results:

First, the computed partial sums s_n all become equal when n is large enough, even though theoretically they should keep increasing to infinity as n increases. This is because in finite precision the new terms we add eventually get so small that they effectively give zero contribution. Thus, although in principle the series diverges, in practice the series appears to converge on the computer. This gives an illustration of idealism vs realism in mathematics!

Second, the backward sum is strictly larger than the forward sum! This is because in the summation a term effectively adds zero when the term is sufficiently small compared to the current partial sum, and the size of the partial sums is vastly different if we add in a forward or backward manner.

37.7 Abel

About Abel, tragic life. Bicentennial. Norwegian hero.

n	forward sum	backward sum
10000	9.787612915039062	9.787604331970214
100000	12.090850830078120	12.090151786804200
1000000	14.357357978820800	14.392651557922360
2000000	15.311032295227050	15.086579322814940
3000000	15.403682708740240	15.491910934448240
5000000	15.403682708740240	16.007854461669920
10000000	15.403682708740240	16.686031341552740
20000000		17.390090942382810
30000000		17.743585586547850
40000000		18.257812500000000
50000000		18.807918548583980
100000000	15.403682708740240	18.807918548583980
200000000		18.807918548583980
1000000000		18.807918548583980

FIGURE 37.2. Forward $1 + \frac{1}{2} + \cdots + \frac{1}{n}$ and backward $\frac{1}{n} + \frac{1}{n-1} + \cdots + \frac{1}{2} + 1$ partial harmonic sums for various n computed with double precision.

Chapter 37 Problems

37.1. Prove that the series $\sum_{i=1}^{\infty} i^{-\alpha}$ converges if and only if $\alpha > 1$. Hint: Compare with a primitive function of $x^{-\alpha}$.

37.2. Prove that the series $\sum_{i=1}^{\infty} (-1)^i i^{-\alpha}$ converges if and only if $\alpha > 0$.

37.3. Prove that the following series converges: (a) $\displaystyle\sum_{i=1}^{\infty} e^{-i}$. (b) $\displaystyle\sum_{i=1}^{\infty} \frac{1+(-1)^i}{i^2}$. (c) $\displaystyle\sum_{i=1}^{\infty} \frac{e^{-i}}{i}$. (d) $\displaystyle\sum_{i=1}^{\infty} \frac{1}{(i+1)(i+4)}$.

37.4. Prove that $\displaystyle\sum_{i=1}^{\infty} \frac{1}{i^2 - i}$ converges. Hint: first show that $\frac{1}{2}i^2 - i \geq 0$ for $i \geq 2$.

37.5. Estimate the remainder $\sum_{i=n}^{\infty} \frac{1}{i^2}$ for different values of n.

37.6. Prove that $\displaystyle\sum_{i=1}^{\infty} (-1)^i \sin(1/i)$ converges. More difficult: prove that it is **not** absolutely convergent.

37.7. Explain in detail why the backward partial sum of the series $\sum_{i=1}^{\infty} \frac{1}{i}$ is larger than the forward sum.

38
Scalar autonomous initial value problems

He doesn' use long, difficult words, like Owl. (The House at Pooh Corner, Milne)

38.1 Introduction

In this chapter, we consider the initial value problem for a *scalar autonomous non-linear differential equation*: Find a function $u : [0,1] \to \mathbb{R}$ such that

$$u'(x) = f(u(x)) \quad \text{for } 0 < x \leq 1, \, u(0) = u_0, \tag{38.1}$$

where $f : \mathbb{R} \to \mathbb{R}$ is a given function and u_0 a given initial value. We assume that $f : \mathbb{R} \to \mathbb{R}$ is bounded and Lipschitz continuous, that is, there are constants L_f and M_f such that for all $v, w \in \mathbb{R}$,

$$|f(v) - f(w)| \leq L_f|v - w|, \quad \text{and} \quad |f(v)| \leq M_f. \tag{38.2}$$

For definiteness, we choose the interval $[0,1]$, and we may of course generalize to any interval $[a,b]$.

The problem (38.1) is in general *non-linear*, since $f(v)$ in general is non-linear in v, that is, $f(u(x))$ depends non-linearly on $u(x)$. We have already in Chapter *The exponential function* considered the basic case with f linear, which is the case $f(u(x)) = u(x)$ or $f(v) = v$. Now we pass on to nonlinear functions such as $f(v) = v^2$ and others.

Further, we call (38.1) *autonomous* because $f(u(x))$ depends on the value of the solution $u(x)$, but not directly on the independent variable x. A *non-autonomous* differential equation has the form $u'(x) = f(u(x), x)$, where

$f(u(x), x)$ depends on both $u(x)$ and x. The differential equation $u'(x) = xu^2(x)$ is non-autonomous and non-linear with $f(v, x) = xv^2$, while the equation $u'(x) = u(x)$ defining the exponential is autonomous and linear with $f(v) = v$.

Finally, we refer to (38.1) as a *scalar* problem since $f : \mathbb{R} \to \mathbb{R}$ is a real valued function of one real variable, that is, $v \in \mathbb{R}$ and $f(v) \in \mathbb{R}$, and thus $u(x)$ takes real values or $u : [0, 1] \to \mathbb{R}$. Below we shall consider *systems* of equations with $f : \mathbb{R}^d \to \mathbb{R}^d$ and $u : [0, 1] \to \mathbb{R}^d$, where $d > 1$, which models a very large range of phenomena.

We hope the reader (like Owl) is now at ease with the terminology: In this chapter we thus focus on scalar autonomous non-linear differential equations.

The initial value problem for a scalar autonomous differential equation is the simplest of all initial value problems and the solution (when it exists) can be expressed analytically in terms of a primitive function $F(v)$ of the function $1/f(v)$. In the next chapter we present an extension of this solution formula to a certain class of scalar non-autonomous differential equations referred to as *separable* differential equations. The analytical solution formula does not generalize to an intial value problems for a system of differential equations, and is thus of very very limited use. However, the solution formula is really a beautiful application of Calculus, which may give valuable information in compact form in the special cases when it is applicable.

We also present a direct constructive proof of existence of a solution to the scalar autonomous problem, which generalizes to the very general case of a initial value problems for (autonomous and non-autonomous) systems of differential equations, as presented in Chapter *The general initial value problem* below.

38.2 An analytical solution formula

To derive the analytical solution formula, we let $F(v)$ be a primitive function of the function $1/f(v)$, assuming v takes values so that zeros of $f(v)$ are avoided. Observe that here $F(v)$ is a primitive function of the function $1/f(v)$, and not of $f(v)$. We can then write the equation $u'(x) = f(u(x))$ as

$$\frac{d}{dx}F(u(x)) = 1,$$

since by the Chain rule $\frac{d}{dx}F(u(x)) = F'(u(x))u'(x) = \frac{u'(x)}{f(u(x))}$. We conclude that

$$F(u(x)) = x + C,$$

where the constant C is to be determined by the initial condition by setting $F(u_0) = C$ at $x = 0$. Formally, we can carry out the calculus as follows:

We write the differential equation $\frac{du}{dx} = f(u)$ in the form

$$\frac{du}{f(u)} = dx$$

and integrate to get

$$F(u) = x + C,$$

which gives the solution formula

$$u(x) = F^{-1}(x + F(u_0)), \tag{38.3}$$

where F^{-1} is the inverse of F.

The model $u' = u^n$ for $n > 1$

We use this example to show that the nonlinear nature of (38.1) allows the interesting behavior of *finite-time-blow-up* of the solution. First consider the case $n = 2$, that is, the initial value problem

$$u'(x) = u^2(x) \quad \text{for } x > 0, \quad u(0) = u_0 > 0, \tag{38.4}$$

with $f(v) = v^2$. In this case $F(v) = -1/v$ with $F^{-1}(w) = -1/w$, and we obtain the solution formula

$$u(x) = \frac{1}{u_0^{-1} - x} = \frac{u_0}{1 - u_0 x}.$$

We see that that $u(x) \to \infty$ as $x \to u_0^{-1}$, that is, the solution $u(x)$ of (38.1) with $f(u) = u^2$ tends to infinity as x increases to u_0^{-1} and the solution does not exist beyond this point, see Fig. 38.1. We say that the solution u *blows up* in finite time or exhibits *finite time blow-up*.

If we consider $u'(x) = u^2(x)$ as a model for the growth of a quantity $u(x)$ with time x in which the rate of growth is proportional to $u^2(x)$ and compare with the model $u'(x) = u(x)$ with solution $u_0 \exp(x)$ showing exponential growth. In the model $u'(x) = u^2(x)$ the growth is eventually much quicker than exponential growth since $u^2(x) > u(x)$ as soon as $u(x) > 1$.

We now generalize to

$$u'(x) = u^n(x) \quad \text{for } x > 0, \quad u(0) = u_0,$$

where $n > 1$. In this case $f(v) = v^{-n}$ and $F(v) = -\frac{1}{n-1} v^{-(n-1)}$, and we find the solution formula

$$u(x) = \frac{1}{(u_0^{-n+1} - (n-1)x)^{1/(n-1)}}.$$

Again the solution exhibits finite time blow-up.

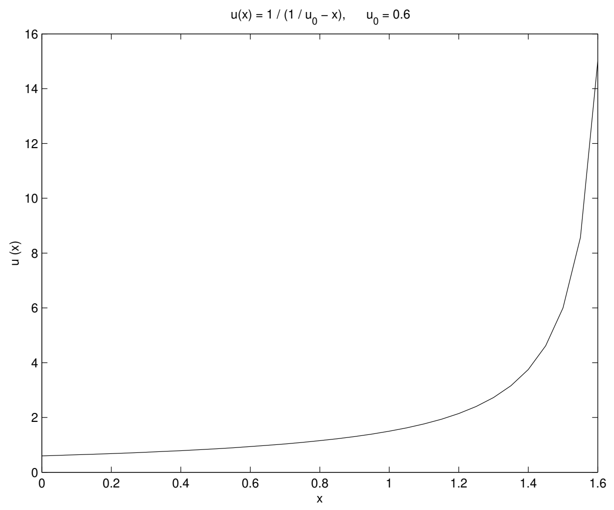


FIGURE 38.1. Solution of the equation $u' = u^2$

The logistic equation $u' = u(1-u)$

We now consider the initial value problem for the *logistic equation*

$$u'(x) = u(x)(1-u(x)) \quad \text{for } x > 0, \quad u(0) = u_0,$$

which was derived by the mathematician and biologist Verhulst as a model of a population with the *growth rate* decreasing with the factor $(1-u)$, as compared with the basic model $u' = u$, as the population approaches the value 1. Typically we assume $0 < u_0 < 1$ and expect to have $0 \leq u(x) \leq 1$.

In this case we have $f(u) = \frac{1}{u(1-u)}$ and using that $f(u) = \frac{1}{u} + \frac{1}{1-u}$, we find that

$$F(u) = \log(u) - \log(1-u) = \log(\frac{u}{1-u}),$$

so that

$$\log(\frac{u}{1-u}) = x + C,$$

or

$$\frac{u}{1-u} = \exp(C)\exp(x).$$

Solving for u and using the initial condition we find that

$$u(x) = \frac{1}{\frac{1-u_0}{u_0}\exp(-x) + 1}.$$

We see that the solution $u(x)$ increases from $u_0 < 1$ to 1 as x increases to infinity, see Fig. 38.2, which gives the famous logistic S-curve modeling growth with decreasing growth rate.

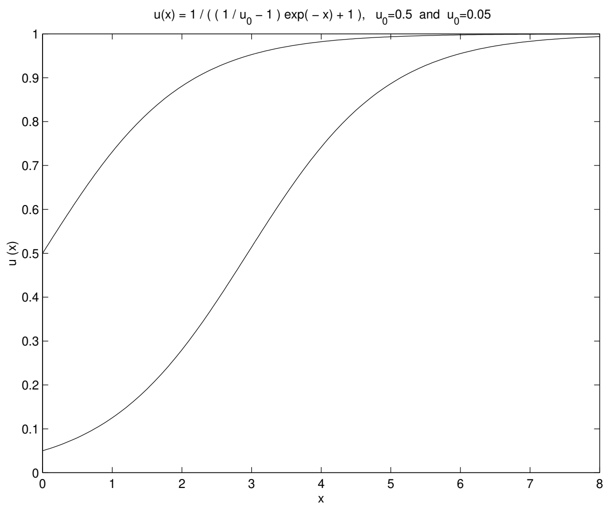


FIGURE 38.2. Solution of the logistic equation

38.3 Construction of the solution

For the direct construction of a solution of (38.1), we shall use the same technique as that used for the linear problem $f(u(x)) = u(x)$ considered in Chapter *The exponential function.* Of course, one may ask why we should worry about constructing the solution, when we already have the solution formula (38.3). We may reply that the solution formula involves the (inverse of) the primitive function $F(v)$ of $1/f(v)$, which we may have to construct anyway, and then a direct construction of the solution may in fact be preferrable. In general, a solution formula when available may give valuable information about qualititativ properties of the solution such as dependence of parameters of the problem, even if it is not necessarily the most effective way of actually computing the solution.

To construct the solution we introduce meshes with nodes $x_i^n = ih_n$ for $i = 1, \cdots, N$, where $h_n = 2^{-n}$ and $N = 2^n$, and for $n = 1, 2, \cdots$. we then define an approximate continuous piecewise linear solution $U^n(x)$ for $0 < x \leq 1$ by the formula

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n f(U^n(x_{i-1}^n)) \quad \text{for } i = 1, \cdots, N, \tag{38.5}$$

with $U^n(0) = u_0$.

We want to prove that $\{U^n(x)\}$ is a Cauchy sequence for $x \in [0, 1]$ and we start by estimating $U^n(x_i^n) - U^{n+1}(x_i^n)$ for $i = 1, \cdots, N$. Taking two steps with step size $h_{n+1} = \frac{1}{2}h_n$ to go from time $x_{i-1}^n = x_{2i-2}^{n+1}$ to $x_i^n = x_{2i}^{n+1}$, we get

$$
\begin{aligned}
U^{n+1}(x_{2i-1}^{n+1}) &= U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-2}^{n+1})),\\
U^{n+1}(x_{2i}^{n+1}) &= U^{n+1}(x_{2i-1}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-1}^{n+1})).
\end{aligned}
$$

Inserting now the value of $U^{n+1}(x_{2i-1}^{n+1})$ at the intermediate step x_{2i-1}^{n+1} from the first equation into the second equation gives

$$
\begin{aligned}
U^{n+1}(x_{2i}^{n+1}) = U^{n+1}(x_{2i-2}^{n+1}) &+ h_{n+1} f(U^{n+1}(x_{2i-2}^{n+1}))\\
&+ h_{n+1} f\big(U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-2}^{n+1}))\big).
\end{aligned} \tag{38.6}
$$

Setting $e_i^n \equiv U^n(x_i^n) - U^{n+1}(x_{2i}^{n+1})$ and subtracting (38.6) from (38.5), we get

$$
\begin{aligned}
e_i^n &= e_{i-1}^n + h_n\big(f(U^n(x_{i-1}^n)) - f(U^{n+1}(x_{2i-2}^{n+1}))\big)\\
&\quad + h_{n+1}\Big(f(U^{n+1}(x_{2i-2}^{n+1})) - f\big(U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-2}^{n+1}))\big)\Big)\\
&\equiv e_{i-1}^n + F_{1,n} + F_{2,n},
\end{aligned}
$$

with the obvious definition of $F_{1,n}$ and $F_{2,n}$. Using the Lipschitz continuity and boundedness (38.2), we have

$$
\begin{aligned}
|F_{1,n}| &\le L_f h_n |e_{i-1}^n|,\\
|F_{2,n}| &\le L_f h_{n+1}^2 |f(U^{n+1}(x_{2i-2}^{n+1}))| \le L_f M_f h_{n+1}^2.
\end{aligned}
$$

Thus for $i = 1, \cdots, 2^N$,

$$|e_i^n| \le (1 + L_f h_n)|e_{i-1}^n| + L_f M_f h_{n+1}^2.$$

Iterating this inequality over i and using that $e_0^n = 0$, we get

$$|e_i^n| \le L_f M_f h_{n+1}^2 \sum_{k=0}^{i-1} (1 + L_f h_n)^k \quad \text{for } i = 1, \cdots, N.$$

Now recalling (31.10) and (31.27), we have

$$\sum_{k=0}^{i-1} (1 + L_f h_n)^k \le \frac{\exp(L_f) - 1}{L_f h_n},$$

and thus we have proved that for $i = 1, \cdots, N$,

$$|e_i^n| \le \frac{1}{2} M_f \exp(L_f) h_{n+1},$$

that is, for $\bar{x} = i h_n$ with $i = 0, ..., N$,

$$|U^n(\bar{x}) - U^{n+1}(\bar{x})| \le \frac{1}{2} M_f \exp(L_f) h_{n+1}.$$

Iterating this inequality as in the proof of the Fundamental Theorem, we get for $m > n$ and $\bar{x} = ih_n$ with $i = 0, ..., N$,

$$|U^n(\bar{x}) - U^m(\bar{x})| \leq \frac{1}{2} M_f \exp(L_f) h_n.$$

Again as in the proof of the Fundamental Theorem, we conclude that $\{U^n(x)\}$ is a Cauchy sequence for each $x \in [0,1]$, and thus converges to a function $u(x)$, which by the construction satisfies the differential equation $u'(x) = f(u(x))$ for $x \in (0,1]$ and $u(0) = u_0$, and thus the limit $u(x)$ is a solution of the initial value problem (38.1).

It remains to prove uniqueness. Assume that $v(x)$ satisfies $v'(x) = f(v(x))$ for $x \in (0,1]$ and $v(0) = u_0$, and consider the function $w = u - v$. Since $w(0) = 0$,

$$\begin{aligned} |w(x)| &= |\int_0^x w'(y)\,dy| = |\int_0^x f(u(y)) - f(v(y))\,dy| \\ &\leq \int_0^x |f(u(y)) - f(v(y))|\,dy \leq \int_0^x L_f |w(y)|\,dy. \end{aligned}$$

Setting $a = \max_{0 \leq x \leq (2L_f)^{-1}} |w(x)|$, we have

$$a \leq \int_0^{(2L_f)^{-1}} L_f a\,dy \leq \frac{1}{2} a$$

which proves that $w(x) = 0$ for $0 \leq x \leq (2L_f)^{-1}$. We now repeat the argument for $x \geq (2L_f)^{-1}$ to get uniqueness for $0 \leq x \leq 1$.

We have now proved:

Theorem 38.1 *The initial value problem* (38.1) *with* $f : \mathbb{R} \to \mathbb{R}$ *Lipschitz continuous and bounded has a unique solution* $u : [0,1] \to \mathbb{R}$, *which is the limit of the sequence of continuous piecewise linear functions* $\{U^n(x)\}$ *constructed from* (38.5) *and satisfying* $|u(x) - U^n(x)| \leq \frac{1}{2} M_f \exp(L_f) h_n$ *for* $x \in [0,1]$.

The attentive reader will note that the existence proof does not seem to apply to e.g. the initial value problem (38.4), because the function $f(v) = v^2$ is not Lipschitz continuous and bounded on $\mathbb{R}$. In fact, the solution $u(x) = \frac{u_0}{1 - u_0 x}$ only exists on the interval $[0, u_0^{-1}]$ and blows up at $x = u_0^{-1}$. However, we can argue that *before* blow-up with say $|u(x)| \leq M$ for some (large) constant M, it suffices to consider the function $f(v) = v^2$ on the interval $[-M, M]$ where the assumption of Lipschitz continuity and boundedness is satisfied. We conclude that for functions $f(v)$ which are Lipschitz continuous and bounded on bounded intervals of $\mathbb{R}$, the constructive existence proof applies as longs as the solution does not blow up.

Chapter 38 Problems

38.1. Solve the following initial value problem analytically: $u'(x) = f(u(x))$ for $x > 0$, $u(0) = u_0$, with (a) $f(u) = -u^2$, (b) $f(u) = \sqrt{u}$, (c) $f(u) = u\log(u)$, (d) $f(u) = 1 + u^2$, (e) $f(u) = \sin(u)$, (f) $f(u) = (1+u)^{-1}$, (g) $f(u) = \sqrt{u^2+4}$.

38.2. Verify that the constructed function $u(x)$ satisfies (38.1). Hint: Use that by the construction we have $u(x) = u_0 + \int_0^x f(u(y))\,dy$ for $x \in [0,1]$.

38.3. Find the velocity of a parachute jumper assuming that the air resistance is proportional to the square of the velocity.

38.4. Let $u(t)$ be the position of a body sliding along x-axis with the velocity $\dot{u}(t)$ satisfying $\dot{u}(t) = -\exp(-u)$. How long time does it take for the body to reach the position $u = 0$ starting from $u(0) = 5$

39
Separable scalar initial value problems

> The search for general methods for integrating ordinary differential equations ended about 1755. (Mathematical Thought, from Ancient to Modern Times, Kline)

39.1 Introduction

We now consider the initial value problem for a scalar non-autonomous differential equation:

$$u'(x) = f(u(x), x) \quad \text{for } 0 < x \leq 1,\ u(0) = u_0, \tag{39.1}$$

in the special case when $f(u(x), x)$ has the form

$$f(u(x), x) = \frac{h(x)}{g(u(x))}, \tag{39.2}$$

where $h : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$. We thus consider the initial value problem

$$u'(x) = \frac{h(x)}{g(u(x))} \quad \text{for } 0 < x \leq 1,\ u(0) = u_0, \tag{39.3}$$

where $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ are given functions, which we refer to as a *separable* problem, because the right hand side $f(u(x), x)$ separates into the quotient of one function $h(x)$ of x only, and one function $g(u(x))$ of $u(x)$ only according to (39.2).

39.2 An analytical solution formula

We shall now derive an analytical solution formula that generalizes the solution formula (38.3) for a scalar autonomous problem (corresponding to the case $h(x) = 1$). Let then $G(v)$ and $H(x)$ be primitive functions of $g(v)$ and $h(x)$ so that $\frac{dG}{dv} = g$ and $\frac{dH}{dx} = h$, and suppose that the function $u(x)$ solves the equation

$$G(u(x)) = H(x) + C, \tag{39.4}$$

for $x \in [0, 1]$, where C is a constant. Differentiating with respect to x using the Chain rule on the left hand side, we then find that $g(u(x))u'(x) = h(x)$, that is $u(x)$ solves the differential equation $u'(x) = h(x)/g(u(x)) = f(u(x), x)$ as desired. Choosing the constant C so that $u(0) = u_0$, we thus obtain a solution $u(x)$ of (39.3), that is the problem (39.1) with $f(u(x), x)$ of the separable form (39.2).

Note that (39.4) is an algebraic equation for the value of the solution $u(x)$ for each value of x. We have thus rewritten the differential equation (39.3) as an algebraic equation (39.4) with x acting as a parameter, and involving primitive functions of $g(y)$ and $h(x)$.

Of course, we may consider (39.1) with x in an interval $[a, b]$ or $[a, \infty)$ with $a, b \in \mathbb{R}$.

Example 39.1. Consider the separable initial value problem

$$u'(x) = xu(x), \quad x > 0,\, u(0) = u_0, \tag{39.5}$$

where $f(u(x), x) = h(x)/g(u(x))$ with $g(v) = 1/v$ and $h(x) = x$. The equation $G(u(x)) = H(x) + C$ takes the form

$$\log(u(x)) = \frac{x^2}{2} + C, \tag{39.6}$$

and thus the solution $u(x)$ of (39.5) is given by the formula

$$u(x) = \exp(\frac{x^2}{2} + C) = u_0 \exp(\frac{x^2}{2}),$$

with $\exp(C) = u_0$ chosen so that the initial condition $u(0) = u_0$ is satisfied. We check by differentiation using the Chain rule that indeed $u_0 \exp(\frac{x^2}{2})$ satisfies $u'(x) = xu(x)$ for $x > 0$.

Formally ("multiplying by dx"), we can rewrite (39.5) as

$$\frac{du}{u} = x\,dx$$

and integrate to get

$$\log(u) = \frac{x^2}{2} + C,$$

which corresponds to the equation (39.6).

EXAMPLE 39.2. On the rainy evening of November 11 1675 Leibniz successfully solved the following problem as a first (crucial) test of the power of the Calculus he had discovered on October 29: Find a curve $y = y(x)$ such that the *subnormal* p, see Fig. 39.1, is inversely proportional to y. Leibniz argued as follows: By similarity, see again Fig. 39.1, we have

$$\frac{dy}{dx} = \frac{p}{y},$$

and assuming the subnormal p to be inversely proportional to y, that is,

$$p = \frac{\alpha}{y}$$

with α a positive constant, we get the differential equation

$$\frac{dy}{dx} = \frac{\alpha}{y^2} = \frac{h(x)}{g(y)}, \tag{39.7}$$

which is separable with $h(x) = \alpha$ and $g(y) = y^2$. The solution $y = y(x)$ with $y(0) = 0$ thus is given by, see Fig. 39.1,

$$\frac{y^3}{3} = \alpha x, \quad \text{that is } y = (3\alpha x)^{\frac{1}{3}}, \tag{39.8}$$

The next morning Leibniz presented his solution to a stunned audience of collegues in Paris, and rocketed to fame as a leading mathematician and Inventor of Calculus.

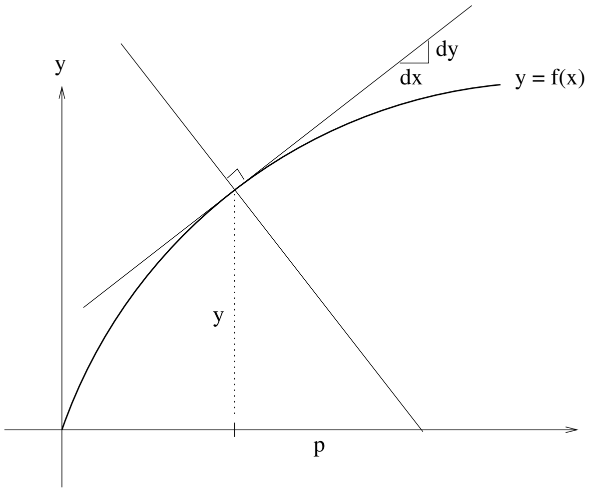


FIGURE 39.1. Leibniz' subnormal problem (change from $y = f(x)$ to $y = y(x)$)

39.3 Volterra-Lotka's predator-prey model

We now consider a biological system consisting of prey and predators like rabbits and foxes which interact. Let $x(t)$ be the density of the prey and $y(t)$ that of the predators at time t and consider *Volterra-Lotka's predator-prey model* for their interaction:

$$\begin{aligned}\dot{x}(t) &= ax(t) - bx(t)y(t), \\ \dot{y}(t) &= -\alpha y(t) + \beta x(t)y(t)\end{aligned} \tag{39.9}$$

where a, b, α and β are positive constants, and $\dot{x} = \frac{dx}{dt}$ and $\dot{y} = \frac{dy}{dt}$. The model includes a growth term $ax(t)$ for the prey corresponding to births and a decay term $bx(t)y(t)$ proportional to the density of prey and predators corresponding to the consumption of prey by the predators, together with corresponding terms for the predators with different signs.

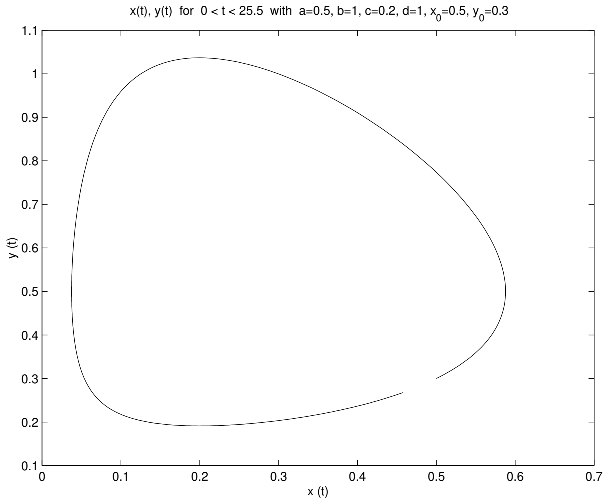


FIGURE 39.2. Phase plane plot of a solution of Volterra-Lotka's equation

This is a system of two differential equations in two unknowns $x(t)$ and $y(t)$ for which analytical solutions are unknown in general. However, we can derive an equation satisfied by the points $(x(t), y(t))$ in an $x - y$ plane, referred to as the $x - y$ *phase plane*, as follows: Dividing the two equations, we get

$$\frac{\dot{y}}{\dot{x}} = \frac{-\alpha y + \beta xy}{ax - bxy}$$

and formally replacing $\frac{\dot{y}}{\dot{x}}$ (by formally dividing out the common dt), we are led to the equation

$$y'(x) = \frac{-\alpha y + \beta xy}{ax - bxy} = \frac{y(-\alpha + \beta x)}{(a - by)x},$$

where $y' = \frac{dy}{dx}$, which is a separable equation with solution $y = y(x)$ satsifying

$$a\log(y) - by = -\alpha\log(x) + \beta x + C,$$

or

$$y^a \exp(-by) = \exp(C)x^{-\alpha}\exp(\beta x)$$

where C is a constant determined by the initial conditions. We plot pairs of (x, y) satisfying this equation in Fig. 39.2 as we let the prey x vary, which traces a *phase plane curve* of the solution $(x(t), y(t)$ of Fig. 39.2 as t varies. We see that the solution is periodic with a variation from (many rabbits, many foxes) to (few rabbits, many foxes) to (few rabbits, few foxes) to (many rabbits, few foxes) and back to (many rabbits, many foxes). Note that the phase plane curve shows the different combinations of rabbits and foxes (x, y), but does *not* give the time evolution $(x(t), y(t))$ of their interaction as a function of time t. We know that for a given t, the point $(x(t), y(t))$ lies on the phase plane curve, but not where.

39.4 A generalization

We now consider a generalization of the separable differential equation (39.3) with solution $u(x)$ satisfying an equation of the form $G(u(x)) - H(x) = C$, to a differential equation with solution satisfying a more general equation of the form $F(x, u(x)) = C$. This closely couples to Chapter *Potential fields* below, and uses a generalization of the Chain rule, which can be accepted right now by a willing reader, and which we will meet again in Chapter *Vector-valued functions of several variables* below.

We thus consider the scalar initial value problem

$$u'(x) = f(u(x), x) \quad \text{for } 0 < x \leq 1,\ u(0) = u_0, \tag{39.10}$$

in the case $f(u(x), x)$ has the form

$$f(u(x), x) = \frac{h(u(x), x)}{g(u(x)), x)}, \tag{39.11}$$

where $h(v, x)$ and $g(v, x)$ are functions of v and x with the special property that

$$g(v, x) = \frac{\partial F}{\partial v}(v, x), \qquad h(v, x) = -\frac{\partial F}{\partial x}(v, x), \tag{39.12}$$

where $F(v, x)$ is a given function of v and x. Above we considered the case when $g(v, x) = g(v)$ is a function of v only and $h(v, x) = h(x)$ is a function of x only, and $F(v, x) = G(v) - H(x)$ with $G(v)$ and $H(x)$ primitive functions of $g(v)$ and $h(x)$, respectively. Now we allow $F(v, x)$ to have a more general form.

Assume now that $u(x)$ satisfies the equation

$$F(u(x), x) = C \qquad \text{for } 0 < x \leq 1.$$

Differentiating both sides with respect to x, using a generalization of the Chain rule, we then get

$$\frac{\partial F}{\partial u}\frac{du}{dx} + \frac{\partial F}{\partial x}\frac{dx}{dx} = g(x, u(x))u'(x) - h(x, u(x)) = 0,$$

and thus $u'(x)$ solves (39.10) with $f(u(x), x)$ of the form (39.11). Again, we thus have rewritten a differential equation as an algebraic equation $F(x, u(x)) = C$ with x acting as a parameter. We give an example. The reader can construct many other similar examples.

EXAMPLE 39.3. Let $F(v, x) = \frac{x^3}{3} + xv + \frac{v^3}{3}$ so that $g(v, x) = \frac{\partial F}{\partial v} = x + v^2$ and $h(v, x) = -\frac{\partial F}{\partial x} = -x^2 - v$. If $u(x)$ satisfies the algebraic equation $\frac{x^3}{3} + xu(x) + \frac{u^3(x)}{3} = C$ for $x \in [0, 1]$, then $u(x)$ solves the differential equation

$$u'(x) = -\frac{x^2 + u(x)}{x + u^2(x)} \qquad \text{for } 0 < x < 1.$$

To sum up: In this chapter we have given analytical solution formula for some special cases of the scalar initial value problem (39.1), but we were not able to give a solution formula in the case of a general non-autonomous scalar equation.

Chapter 39 Problems

39.1. Prove that solutions $(x(t), y(t)$ of the Volterra-Lotka model satisfies

$$\bar{x} = \frac{1}{T}\int_0^T x(t)\,dt = \frac{c}{d}, \quad \bar{y} = \frac{1}{T}\int_0^T y(t)\,dt = \frac{a}{b},$$

where T is the period of periodic solutions. Investigate the effect on the mean values $\bar{x}$ and $\bar{y}$ of hunting of both predator and prey corresponding to including dissipative terms $-\epsilon x$ and $-\epsilon y$ with $\epsilon > 0$. Hint: Consider the integral of $\dot{x}/x$ over a period.

39.2. Extend the Volterra-Lotka model to the model

$$\begin{aligned} \dot{x}(t) &= ax(t) - bx(t)y(t) - ex^2(t), \\ \dot{y}(t) &= -cy(t) + dx(t)y(t) - fy^2(t), \end{aligned} \tag{39.13}$$

where e and f are positive constants, with the additional terms modeling negative influences from competition within the species as the populations densities increase. Compare the solutions of the two models numerically. Is the extended system separable?

39.3. Consider the spread of an infection modeled by

$$\dot{u} = -auv,$$
$$\dot{v} = auv - bv,$$

where $u(t)$ is the density of the susceptibles and $v(t)$ is that of the infectives at time t, and a and b are positive constants. The term $\pm auv$ models the transfer of susceptibles to infectives at a rate proportional to auv, and $-bv$ models the decay of infectives by death or immunity. Study the qualitative behaviour of phase plane curves.

39.4. Extend the previous model by changing the first equation to $\dot{u} = -auv + \mu$, with μ a positive constant modeling a constant growth of the susceptibles. Find the equilibrium point, and study the linearized model linearized at the equilibrium point.

39.5. Motivate the following model for a national economy:

$$\dot{u} = u - av, \quad \dot{v} = b(u - v - w),$$

where u is the national income, v the rate of consumer spending and w the rate of government spending, and $a > 0$ and $b \geq 1$ are constants. Show that if w is constant, then there is an equilibrium state, that is a solution independent of time satisfying $u - av = b(u - v - w) = 0$. Show that the economy oscillates if $b = 1$. Study the stability of solutions. Study a model with $w = w_0 + cu$ with w_0 a constant. Show that there is no equilibrium state in this model if $c \geq (a-1)/a$. Draw some conclusion. Study a model with $w = w_0 + cu^2$.

39.6. Consider a boat being rowed across a river occupying the strip $\{(x, y) : 0 \leq x \leq 1, y \in \mathbb{R}\}$, in such a way that the boat always points in the direction of $(0, 0)$. Assume that the boat moves with the constant speed u relative to the water and that the river flows with constant speed v in the positive y-direction. Show that the equations of motion are

$$\dot{x} = -\frac{ux}{\sqrt{x^2 + y^2}}, \quad \dot{y} = -\frac{uy}{\sqrt{x^2 + y^2}}.$$

Show that the phase-plane curves are given by

$$y = \sqrt{x^2 + y^2} = Ax^{1-\alpha}, \quad \text{where } \alpha - \frac{v}{u}.$$

What happens if $v > u$?. Compute solutions.

40
The general initial value problem

Things are as they are because they were as they were. (Thomas Gold)

40.1 Introduction

We now consider the Initial Value Problem or IVP for a *system* of nonlinear differential equations of the form: Find $u : [0,1] \to \mathbb{R}^d$ such that

$$u'(x) = f(u(x), x) \quad \text{for } 0 < x \leq 1,\, u(0) = u^0, \tag{40.1}$$

where $f : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ is a given bounded and Lipschitz continuous function, $u^0 \in \mathbb{R}^d$ is a given initial value, and $d \geq 1$ is the dimension of the system. The reader may assume $d = 2$ or $d = 3$, recalling the chapters on analytic geometry in $\mathbb{R}^2$ and $\mathbb{R}^3$, and extend to the case $d > 3$ after having read the chapter on analytic geometry in $\mathbb{R}^n$ below. The material in Chapter *Vector-valued functions of several real variables* is largely motivated from the need of studying problems of the form (40.1), and there is thus a close connection between this chapter and the present one. We keep this chapter abstract (and a bit philosophical), and present many concrete examples below.

The IVP (40.1) is the non-autonoumuos vector version of the scalar initial value problem (38.1), and reads as follows in component form: Find functions $u_i : [0,1] \to \mathbb{R}$, $i = 1, ..., d$, such that

$$\begin{aligned} u_1'(x) &= f_1(u_1(x), u_2(x), ..., u_d(x), x) \quad \text{for } 0 < x \le 1, \\ u_2'(x) &= f_2(u_1(x), u_2(x), ..., u_d(x), x) \quad \text{for } 0 < x \le 1, \\ &\dots\dots\dots \\ u_d'(x) &= f_d(u_1(x), u_2(x), ..., u_d(x), x) \quad \text{for } 0 < x \le 1, \\ u_1(0) &= u_{10},\ u_2(0) = u_{20},\ u_d(0) = u_d^0, \end{aligned} \tag{40.2}$$

where $f_i : \mathbb{R}^d \times [0,1] \to \mathbb{R}$, $i = 1, ..., d$, are given functions and u_i^0, $i = 1, ..., d$, are given initial values. With vector notation writing $u = (u_1, ..., u_d)$, $f = (f_1, ..., f_d)$ and $u^0 = (u_1^0, ..., u_d^0)$, we may write (40.2) in the compact form (40.1). Of course, writing $f : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$, means that for each vector $v = (v_1, ..., v_d) \in \mathbb{R}^d$ and $x \in [0,1]$ there is assigned a vector $f(v, x) = (f_1(v, x), ..., f_d(v, x)) \in \mathbb{R}^d$, where $f_i(v, x) = f_i(v_1, ..., v_d, x)$.

We assume Lipschitz continuity and boundedness of $f : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ in the form: There are constants L_f and M_f such that for all $v, w \in \mathbb{R}^d$ and $x, y \in [0,1]$,

$$|f(v,x) - f(w,y)| \le L_f(|v - w| + |x - y|) \quad \text{and} \quad |f(v,x)| \le M_f, \tag{40.3}$$

where $|v| = (\sum_{i=1}^d v_i^2)^{1/2}$ is the Euclidean norm of $v = (v_1, ..., v_d) \in \mathbb{R}^d$.

In short, everything looks the same as in the scalar case of (38.1) with the natural extension to a non-autonomous problem, but the vector interpretation makes the actual content of this chapter vastly different from that of Chapter *Scalar autonomous initial value problems*. In particular, there is in general no analytical solution formula if $d > 1$, since the solution formula for $d = 1$ based on the existence of a primitive function of $1/f(v)$, does not generalize to $d > 1$.

We prove the existence of a unique solution of the IVP (40.1) by using a constructive process which is a direct generalization of the method used for the scalar problem (38.1), which was a direct generalization of method used to construct the integral. The result of this chapter is definitely one of the highlights of mathematics (or at least of this book), because of its generality and simplicity: $f : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ can be *any* bounded Lipschitz continuous function with the dimension d arbitrarily large, and the proof looks exactly the same as in the scalar case. Therefore this chapter has a central role in the book and couples closely to several other chapters below including *Analytic geometry in* $\mathbb{R}^n$, *Solving systems of linear equations*, *Linearization and stability of IVP*, *Adaptive IVP-solvers*, *Vector-valued functions of several real variales* and various chapters on applications including mechanical systems, electrical circuits, chemical reactions, and other phenomena. This means that a full appreciation of this chapter can only be made after digesting all this material. Nevertheless, it should be possible to go through this chapter and understand that the general IVP (40.1) can be solved through a constructive process requiring more or less work. This chapter also may be used as a basis for a bit of philosophical discussion on constructive aspects of the World, as we now proceed to do (for the interested reader).

40.2 Determinism and materialism

Before we present the existence proof (which we thus have already seen), we pause to reflect a little on the related *mechanistic/deterministic* view of science and philosophy going back to Descartes and Newton and forming the basis the industrial society leading into our own time. With this view the World is like a big mechanical Clock governed by laws of mechanics, which may be modeled as an initial value problem of the form (40.1) with a certain function f and initial value u^0 at time $x = 0$. The state of this system for positive time is, according to the existence proof, uniquely determined by the function f and u^0, which would support a *deterministic* or *materialistic* view of the World including the mental processes in human beings: everything that will happen in the future is in principle determined by the present state (assuming no blow-up). Of course, this view is in serious conflict with massive everyday experience of unpredictability and our firm belief in the existence of a *free will*, and considerable efforts have gone into resolving this paradox through the centuries without complete success.

Let's see if we can approach this paradox from a mathematical point of view. After all, the determinitic/materialistic view is founded on a proof of existence of a unique solution of an initial value problem of the form (40.1), and thus the roots of the paradox may be hidden in the mathematical proof itself. We will argue that the resolution of the paradox must be coupled to aspects of *predictability* and *computability* of the problem (40.1), which we will now briefly touch upon and return to in more detail below. We hope the reader is open for this type of discussion, seldom met in a Calculus text. We try to point to the necessity of a proper understanding of a mathematical result, which may appear to be very simple and clear, like the existence proof to be presented, but which in fact may require a lot of explanation and qualification to avoid misunderstanding.

40.3 Predictability and computability

The *predictability* of the problem (40.1) concerns the *sensitivity* of the solution to the given *data*, that is, the function f and the initial value u^0. The sensitivity is a measure of the change of the solution under changes of the data f and u^0. If the solution changes very much even for very small changes of data, then the sensitivity is very high. In such a case we need to know the data with very high precision to accurately predict the solution.

We shall see below that solutions of certain initial value problems are highly sensitive to changes in data and in these problems accurate prediction will be impossible. An example is given by the familiar process of tossing a coin, which can be modeled as an initial value problem. In principle, by repeatedly choosing the same initial value, the person tossing the coin should be able to always get heads, for example. However, we all know that this is impossible in practice, because the process is too sensitive to small changes in the initial value (and the corresponding function f). To handle this type of unpredictability the scientific field of *statistics* has been developed.

Similarly, the *computability* of the problem (40.1) concerns (i) the sensitivity of the solution to errors made in constructing the solution according to the existence proof, and (ii) the amount of computational work needed to construct the solution. Usually, (i) and (ii) go hand in hand: if the sensitivity is high, then a lot of work is required and of course the work also increases with the dimension d. A highly sensitive problem with d very large is thus a computational night-mare. To construct the solution of the initial value problem for even a small part of the Universe will thus be practically impossible with any kind of computer, and claiming that in principle the solution is determined would make little sense.

We will meet this problem with painstaking evidence when we turn into numerical methods. We will see that most systems of the form (40.1) with d small ($d \leq 10$ say) may be solved within fractions of a second on a PC, while some systems (like the famous Lorenz system with $d = 3$ to be studied below) quickly will exhaust even supercomputers because of very high sensitivity. We will further see that many systems of practical interest with d large ($d \approx 10^6 - 10^7$) can be solved within minutes/hours on a PC, while accurate modeling of e.g. turbulent flow requires $d \geq 10^{10}$ and super-computer power. The most powerful super-computer in sight, the *Blue Gene* consisting of 10^6 connected PCs to appear in a couple of years, is designed for initial value problems of molecular dynamics of protein folding for the purpose of medical drug design. A landmark in computing was set in 1997 when the chess computer *Deep Blue* put the the world-champion Gary Kasparov chess mate.

The computational work required to solve (40.1) may thus vary considerably. Below we shall successively uncover a bit of this mystery and identify basic features of problems requiring different amounts of computional work.

We will return to the concepts of predictability and computabilty of differential equations below. Here we just wanted to give some perspective on the constructive existence proof to be given showing some limits of mathematics as a human activity.

40.4 Construction of the solution

The construction of the solution $u(x)$ of (40.1) looks identical to the construction of the solution of (38.1), after we interpret $u(x)$ and $f(u(x))$ as vectors instead of scalars and make the natural extension to a non-autonomous problem.

We begin by discretizing $[0,1]$ using a mesh with nodes $x_i^n = ih_n$ for $i = 1, \cdots, N$, where $h_n = 2^{-n}$ and $N = 2^n$. For $n = 1, \cdots, N$, we define an approximate piecewise linear solution $U^n : [0,1] \to \mathbb{R}^d$ by the formula

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n f(U^n(x_{i-1}^n), x_{i-1}^n), \quad \text{for } i = 1, \cdots, N, \tag{40.4}$$

and setting $U^n(0) = u^0$. Note that $U^n(x)$ is linear on each subinterval $[x_{n-1}^n, x_i^n]$.

We want to prove that for $x \in [0,1]$, $\{U^n(x)\}_{n=1}^{\infty}$ is a Cauchy sequence in $\mathbb{R}^d$. We start by estimating $U^n(x_i^n) - U^{n+1}(x_i^n)$ for $i = 1, \cdots, N$. Taking two steps with step size $h_{n+1} = \frac{1}{2}h_n$ to go from time $x_{i-1}^n = x_{2i-2}^{n+1}$ to $x_i^n = x_{2i}^{n+1}$, we have

$$\begin{aligned} U^{n+1}(x_{2i-1}^{n+1}) &= U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n), \\ U^{n+1}(x_{2i}^{n+1}) &= U^{n+1}(x_{2i-1}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-1}^{n+1}), x_{2i-1}^{n+1}). \end{aligned}$$

Inserting now the value of $U^{n+1}(x_{2i-1}^{n+1})$ at the intermediate step x_{2i-1}^{n+1} from the first equation into the second equation, we get

$$\begin{aligned} U^{n+1}(x_{2i}^{n+1}) &= U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n) \\ &\quad + h_{n+1} f\big(U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n), x_{2i-1}^{n+1}\big). \end{aligned} \tag{40.5}$$

Setting $e_i^n \equiv U^n(x_i^n) - U^{n+1}(x_{2i}^{n+1})$ and subtracting (40.5) from (40.4) gives

$$\begin{aligned} e_i^n &= e_{i-1}^n + h_n\big(f(U^n(x_{i-1}^n), x_{i-1}^n) - f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n)\big) \\ &+ h_{n+1}\Big(f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n) - f\big(U^{n+1}(x_{2i-2}^{n+1}) + h_{n+1} f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n), x_{2i-1}^{n+1}\big)\Big) \\ &\equiv e_{i-1}^n + F_{1,n} + F_{2,n}, \end{aligned}$$

with the obvious definitions of $F_{1,n}$ and $F_{2,n}$. Using (40.3), we have

$$\begin{aligned} |F_{1,n}| &\le L_f h_n |e_{i-1}^n|, \\ |F_{2,n}| &\le L_f h_{n+1}^2 (|f(U^{n+1}(x_{2i-2}^{n+1}), x_{i-1}^n)| + 1) \le L_f \bar{M}_f h_{n+1}^2, \end{aligned}$$

where $\bar{M}_f = M_f + 1$, and so for $i = 1, \cdots, N$,

$$|e_i^n| \le (1 + L_f h_n)|e_{i-1}^n| + L_f \bar{M}_f h_{n+1}^2.$$

Iterating this inequality over i and using that $e_0^n = 0$, we get

$$|e_i^n| \le L_f \bar{M}_f h_{n+1}^2 \sum_{k=0}^{i-1} (1 + L_f h_n)^i \quad \text{for } i = 1, \cdots, N.$$

Recalling (31.10) and (31.27), we have

$$\sum_{k=0}^{i-1}(1+L_f h_n)^k = \frac{(1+L_f h_n)^i - 1}{L_f h_n} \leq \frac{\exp(L_f) - 1}{L_f h_n},$$

and thus we have proved that for $i = 1, \cdots, N$,

$$|e_i^n| \leq \frac{1}{2}\bar{M}_f \exp(L_f) h_{n+1},$$

that is, for $\bar{x} = ih_n$ with $i = 0, ..., N$,

$$|U^n(\bar{x}) - U^{n+1}(\bar{x})| \leq \frac{1}{2}\bar{M}_f \exp(L_f) h_{n+1}.$$

Iterating this inequality as in the proof of the Fundamental Theorem, we get for $m > n$ and $\bar{x} = ih_n$ with $i = 0, ..., N$,

$$|U^n(\bar{x}) - U^m(\bar{x})| \leq \frac{1}{2}\bar{M}_f \exp(L_f) h_n. \tag{40.6}$$

Again as in the proof of the Fundamental Theorem, we conclude that $\{U^n(x)\}$ is a Cauchy sequence for each $x \in [0,1]$, and thus converges to a function $u(x)$, which by the construction satisfies the differential equation $u'(x) = f(u(x))$ for $x \in (0,1]$ and $u(0) = u^0$, and thus the limit $u(x)$ is a solution of the initial value problem (40.1). Uniqueness of a solution follows as in the scalar case considered in Chapter *Scalar autonomous initial value problems*. We have now proved the following basic result:

Theorem 40.1 *The initial value problem* (40.1) *with* $f : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ *bounded and Lipschitz continuous, has a unique solution* $u(x)$, *which is the limit of the sequence of continuous piecewise linear functions* $\{U^n(x)\}$ *constructed from* (40.4) *and satisfying*

$$|u(x) - U^n(x)| \leq (M_f + 1)\exp(L_f) h_n \qquad \textit{for } x \in [0,1]. \tag{40.7}$$

40.5 Computational work

The convergence estimate (40.7) indicates that the work required to compute a solution $u(x)$ of (40.1) to a given accuracy is proportional to $\exp(L_f)$ and to $\exp(L_f T)$ if we consider a time interval $[0,T]$ instead of $[0,1]$. With $L_f = 10$ and $T = 10$, which would seem to be a very innocent case, we would have $\exp(L_f T) = \exp(10^2)$ and we would thus have to choose h_n smaller than $\exp(-10^2) \approx 10^{-30}$, and the number of computational operations would be of the order 10^{30} which would be at the limit of any practical possibility. Already moderately large constants such as $L_f = 100$

and $T = 100$, would give an exponential factor $\exp(10^4)$ way beyond any comprehension. We conclude that the appearance of the exponential factor $\exp(L_f T)$, which corresponds to a worst possible case, seems to limit the interest of the existence proof. Of course, the worst possible case does not necessarily have to occur always. Below we will present problems with special features for which the error is actually smaller than worst possible, including the important class of *stiff problems* where large Lipschitz constants cause quick exponential decay instead of expeonential growth, and the Lorenz system where the error growth turns out to be of order $\exp(T)$ instead of $\exp(L_f T)$ with $L_f = 100$.

40.6 Extension to second order initial value problems

Consider a second order initial value problem

$$\ddot{v}(t) = g(v(t), \dot{v}(t)) \text{ for } 0 < t \leq 1,\ v(0) = v_0,\ \dot{v}(0) = \dot{v}_0, \tag{40.8}$$

with initial conditions for $v(0)$ and $\dot{v}(0)$, where $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is Lipschitz continuous, $v : [0, 1] \to \mathbb{R}^d$ and $\dot{v} = \frac{dv}{dt}$. In mechanics, initial value problems often come in such second order form as they express Newton's Law with $\ddot{v}(t)$ representing acceleration and $g(v(t), \dot{v}(t))$ force. This problem can be reduced to a first order system of the form (40.1) by introducing the new variable $w(t) = \dot{v}(t)$ and writing (40.8) as

$$\begin{aligned} &\dot{w}(t) = g(v(t), w(t)) \quad \text{for } 0 < t \leq 1, \\ &\dot{v}(t) = w(t) \quad \text{for } 0 < t \leq 1, \\ &v(0) = v_0,\ w(0) = \dot{v}_0. \end{aligned} \tag{40.9}$$

Setting $u = (u_1, ..., u_{2d}) = (v_1, ..., v_d, w_1, ..., w_d)$ and $f(u) = (g_1(u), ..., g_d(u), u_{d+1}, ..., u_{2d})$, the system (40.9) takes the form $\dot{u}(t) = f(u(t))$ for $0 < t \leq 1$, and $u(0) = (v_0, \dot{v}_0)$.

In particular, we can rewrite the second order scalar equations $\ddot{v} + v = 0$ as a first order system and obtain existence of the trigonometric functions via the general existence result for first order systems as solutions of the corresponding initial value problem with appropriate data.

40.7 Numerical methods

The computational solution of differential equations is an important subject with many aspects. The overall objective may be viewed to be to compute approximate solutions with as little work as possible per digit of accuracy.

So far we have discussed only the simplest method for constructing approximate solutions. In this section, we give a brief glimpse of other methods. In Chapter *Adaptive IVP solvers*, we continue this study.

The computational method we have used so far, in which

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n f(U^n(x_{i-1}^n), x_{i-1}^n), \quad \text{for } i = 1, \cdots, N, \tag{40.10}$$

with $U^n(0) = u^0$, is called the *forward Euler* method. The forward Euler method is an *explicit* method because we can directly compute $U^n(x_i^n)$ from $U^n(x_{i-1}^n)$ without solving a system of equations.

In contrast, the *backward Euler* method in which the approximate solution is computed via the equation

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n f(U^n(x_i^n), x_i^n), \quad \text{for } i = 1, \cdots, N, \tag{40.11}$$

with $U^n(0) = u^0$, is an *implicit* method. At each step we need to solve the system

$$V = U^n(x_{i-1}^n) + h_n f(V, x_i^n), \tag{40.12}$$

to compute $U^n(x_i^n)$ from $U^n(x_{i-1}^n)$. Another implicit method is the *midpoint method*

$$U^n(x_i^n) = U^n(x_{i-1}^n) + h_n f(\frac{1}{2}(U^n(x_{i-1}^n) + U^n(x_i^n)), \bar{x}_{i-1}^n), \quad \text{for } i = 1, \cdots, N, \tag{40.13}$$

with $\bar{x}_{i-1}^n = \frac{1}{2}(x_{i-1}^n + x_i^n)$, where we have to solve the system

$$V = U^n(x_{i-1}^n) + h_n f(\frac{1}{2}(U^n(x_{i-1}^n) + V), \bar{x}_{i-1}^n) \tag{40.14}$$

at each step. Note that both (40.12) and (40.14) are nonlinear equations when f is nonlinear. We may use Fixed Point Iteration or Newton's method to solve them, see Chapter *Vector-valued functions of several real variables* below.

We also present the following variant of the midpoint method, which we call the cG(1), *continuous Galerkin method with trial functions of order* 1:. The approximate solution is computed via

$$U^n(x_i^n) = U^n(x_{i-1}^n) + \int_{x_{i-1}^n}^{x_i^n} f(U(x), x)\, dx, \quad i = 1, \cdots, N, \tag{40.15}$$

and $U^n(0) = u^0$, where $U^n(x)$ is continuous piecewise linear function with the values $U^n(x_i^n)$ at the nodes x_i^n. If we evaluate the integral in (40.15) with the midpoint quadrature rule, we obtain the midpoint method. We can of course use other quadrature formulas to get different methods.

We shall see that cG(1) is the first in a family of methods cG(q) with $q = 1, 2, ...$, where the solution is approximated by continuous piecewise polynomials of order q. The Galerkin feature of cG(1) is expressed by the fact that the method can be formulated as

$$\int_{x_{i-1}^n}^{x_i^n} (\frac{dU^n}{dx}(x) - f(U^n(x), x))\, dx = 0,$$

stating that the mean-value over each subinterval of the *residual* $\frac{dU^n}{dx}(x) - f(U(x), x)$ of the continuous piecewise linear approximate solution $U^n(x)$, is equal to zero (or that the residual is orthogonal to the set of constant functions on each subinterval with a terminology to be used below).

We can prove convergence of the backward Euler and midpoint methods in the same way as for the forward Euler method. The forward and backward Euler methods are *first order accurate* methods in the sence that the error $|u(x) - U^n(x)|$ is proportional to the step size h_n, while the midpoint method is *second order accurate* with the error proportional to h_n^2 and thus in general is more accurate. The computational work per step is generally smaller for an explicit method than for an implicit method, since no system of equations has to be solved at each step. For so-called stiff problems, explicit methods may require very small time steps compared to implicit methods, and then implicit methods can give a smaller total cost. We will return to these issues in Chapter *Adaptive IVP solvers* below.

Note that all of the methods discussed so far generalize to allow non-uniform meshes $0 = x_0 < x_1 < x_2 < ... < x_N = 1$ with possibly varying steps $x_i - x_{i-1}$. We will below return to the problem of *automatic step-size control* with the purpose of keeping the error $|u(x_i) - U(x_i)| \leq TOL$ for $i = 1, \cdots, N$, where TOL is a given tolerance, while using as few time steps as possible by varying the mesh steps, cf. the Chapter *Numerical Quadrature.*

Chapter 40 Problems

40.1. Prove existence of a solution of the initial value problem (40.1) using the backward Euler method or the midpoint method.

40.2. Complete the proof of existence for (40.1) by proving that the constructed limit function $u(x)$ solves the initial value problem. Hint: use that $u_i(x) = \int_0^x f_i(u(y))\, dy$ for $x \in [0, 1]$, $i = 1, ..., d$.

40.3. Give examples of problems of the form (40.1).

41
Calculus Tool Bag I

After experience had taught me that all the usual surroundings of social life are vain and futile; seeing that none of the objects of my fears contained in themselves anything either good or bad, except in so far as the mind is affected by them, I finally resolved to inquire whether there might be some real good having power to communicate itself, which would affect the mind singly, to the exclusion of all else: whether, in fact, there might be anything of which the discovery and attainment would enable me to enjoy continuous, supreme, and unending happiness. (Spinoza)

Sapiens nihil affirmat quod non probat.

41.1 Introduction

We present a *Calculus Tool Bag I* containing a minimal set of important tools and concepts of Calculus for functions $f:\mathbb{R}\to\mathbb{R}$. Below, we present a *Calculus Tool Bag II* containing the corresponding of tools and concepts of Calculus for functions $f:\mathbb{R}^m\to\mathbb{R}^n$.

41.2 Rational numbers

We start with the set of integers $\mathbb{Z}=\{\cdots,-3,-2,-1,0,1,2,3,\cdots\}$ together with the usual operations of addition, subtraction and multiplication. We define the set of rational numbers $\mathbb{Q}$ as the set of pairs (p,q) with

p and $q \neq 0$ integers, and we write $(p, q) = \frac{p}{q}$ along with the arithmetic operations of addition,

$$\frac{p}{q} + \frac{r}{s} = \frac{ps + qr}{qs},$$

multiplication,

$$\frac{p}{q} \times \frac{r}{s} = \frac{pr}{qs},$$

and division

$$(p, q)/(r, s) = \frac{(p, q)}{(r, s)} = (ps, qr),$$

assuming $r \neq 0$. With the operation of division, we can solve the equation $ax = b$ to get $x = b/a$ for $a, b \in \mathbb{Q}$ with $a \neq 0$.

Rational numbers have periodic decimal expansions. There is no rational number x such that $x^2 = 2$.

41.3 Real numbers. Sequences and limits

Definitions: A real number is specified by an *infinite decimal expansion* of the form

$$\pm p_m \cdots p_0.q_1 q_2 q_3 \cdots$$

with a never ending list of decimals q_1, q_2, ..., where each of the p_i and q_j are one of the 10 digits $0, 1, ..., 9$. The set of (all possible) real numbers is denoted by $\mathbb{R}$.

A sequence $\{x_i\}_{i=1}^\infty$ of real numbers converges to a real number x if for any $\epsilon > 0$ there is a natural number N such that $|x_i - x| < \epsilon$ for $i \geq N$ and we write $x = \lim_{i\to\infty} x_i$.

A sequence $\{x_i\}_{i=1}^\infty$ of real numbers is a Cauchy sequence if for all $\epsilon > 0$ there is a natural number N such that

$$|x_i - x_j| \leq \epsilon \quad \text{for} \quad i, j \geq N.$$

Basic properties: A convergent sequence of real numbers is a Cauchy sequence. A Cauchy sequence of real numbers converges to a unique real number. We have $\lim_{i\to\infty} x_i = x$, where $\{x_i\}_{i=1}^\infty$ is the sequence of truncated decimal expansions of x. .

41.4 Polynomials and rational functions

A polynomial function $f : \mathbb{R} \to \mathbb{R}$ of degree n has the form $f(x) = a_0 + a_1 x + \cdots + a_n x^n$ with coefficients $a_i \in \mathbb{R}$. A rational function $h(x)$ has the form $h(x) = f(x)/g(x)$, where $f(x)$ and $g(x)$ are polynomials.

41.5 Lipschitz continuity

Definition: A function $f : I \to \mathbb{R}$, where I is an interval of real numbers, is Lipschitz continuous on I with Lipschitz constant $L_f \geq 0$ if

$$|f(x_1) - f(x_2)| \leq L_f|x_1 - x_2| \quad \text{for all } \ x_1, x_2 \in I.$$

Basic facts: Polynomial functions are Lipschitz continuous on bounded intervals. Sums, products and composition of Lipschitz continuous functions are Lipschitz continuous. Quotients of Lipschitz continuous functions are Lipschitz continuous on intervals where the denominator is bounded away from zero. A Lipschitz continuous function $f : I \to \mathbb{R}$, where I is an interval of real numbers, satisfies:

$$f(\lim_{i\to\infty} x_i) = \lim_{i\to\infty} f(x_i), \tag{41.1}$$

for any convergent sequence $\{x_i\}$ in I with $\lim_{i\to\infty} x_i \in I$.

41.6 Derivative

Definition:The function $f : (a, b) \to \mathbb{R}$ is differentiable at $\bar{x} \in (a, b)$ with derivative $f'(\bar{x}) = \frac{df}{dx}(\bar{x})$ if there are real numbers $f'(\bar{x})$ and $K_f(\bar{x})$ such that for $x \in (a, b)$ close to $\bar{x}$,

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + E_f(x, \bar{x}),$$
$$\text{with } |E_f(x, \bar{x})| \leq K_f(\bar{x})|x - \bar{x}|^2.$$

If the constant $K_f(\bar{x})$ can be chosen independently of $\bar{x} \in (a, b)$, then $f : (a, b) \to \mathbb{R}$ is said to be uniformly differentiable on (a, b).

Derivative of x^α with $\alpha \neq 0$: The derivative of $f(x) = x^\alpha$ is $f'(x) = \alpha x^{\alpha-1}$ for $\alpha \neq 0$, and $x \neq 0$ for $\alpha < 1$.

Bounded derivative implies Lipschitz continuity: If $f(x)$ is uniformly differentiable on the interval $I = (a, b)$ and there is a constant L such that

$$|f'(x)| \leq L, \quad \text{for } x \in I,$$

then $f(x)$ is Lipschitz continuous on I with Lipschitz constant L.

41.7 Differentiation rules

Linear Combination rule:

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

where c is a constant.

Product rule:

$$(fg)'(x) = f(x)g'(x) + f'(x)g(x).$$

Chain rule:

$$(f \circ g)'(x) = f'(g(x))g'(x), \quad \text{or}$$
$$\frac{dh}{dx} = \frac{df}{dy}\frac{dy}{dx},$$

where $h(x) = f(y)$ and $y = g(x)$, that is $h(x) = f(g(x)) = (f \circ g)(x)$.

Quotient rule:

$$\left(\frac{f}{g}\right)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2},$$

provided $g(x) \neq 0$.

The derivative of an inverse function:

$$\frac{d}{dy}f^{-1}(y) = \frac{1}{\frac{d}{dx}f(x)}.$$

where $y = f(x)$ and $x = f^{-1}(y)$.

41.8 Solving $f(x) = 0$ with $f : \mathbb{R} \to \mathbb{R}$

Bisection: If $f : [a, b] \to \mathbb{R}$ is Lipschitz continuous on $[a, b]$ and $f(a)f(b) < 0$, then the Bisection algorithm converges to a root $\bar{x} \in [a, b]$ of $f(x) = 0$.

Fixed Point Iteration: A Lipschitz continuous function $g : \mathbb{R} \to \mathbb{R}$ with Lipschitz constant $L < 1$ is said to be a contraction mapping. A contraction mapping $g : \mathbb{R} \to \mathbb{R}$ has a unique fixed point $\bar{x} \in \mathbb{R}$ satisfying $\bar{x} = g(\bar{x})$ and any sequence $\{x_i\}_{i=1}^{\infty}$ generated by Fixed Point Iteration $x_i = g(x_{i-1})$ converges to $\bar{x}$.

Newton's method: Newton's method $x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}$ for computing a root $\bar{x}$ of $f : \mathbb{R} \to \mathbb{R}$ converges quadratically if $f'(x)$ is bounded away from zero for x close to $\bar{x}$ and the initial approximation is sufficiently close to the root $\bar{x}$.

41.9 Integrals

The Fundamental Theorem of Calclulus: If $f : [a, b]$ is Lipschitz continuous, then there is a unique uniformly differentiable function $u : [a, b] \to$

$\mathbb{R}$, that solves the initial value problem

$$\begin{cases} u'(x) = f(x) & \text{for } x \in (a, b], \\ u(a) = u_a, \end{cases}$$

where $u_a \in \mathbb{R}$ is given. The function $u : [a, b] \to \mathbb{R}$ can be expressed as

$$u(\bar{x}) = u_a + \int_a^{\bar{x}} f(x)\,dx \quad \text{for } \bar{x} \in [a, b],$$

where

$$\int_0^{\bar{x}} f(x)\,dx = \lim_{n\to\infty} \sum_{i=1}^{j} f(x_{i-1}^n) h_n,$$

with $\bar{x} = x_j^n$, $x_i^n = a + ih_n$, $h_n = 2^{-n}(b-a)$. More precisely, if the Lipschitz constant of f is L_f then for $n = 1, 2, ...,$

$$\Big| \int_a^{\bar{x}} f(x)\,dx - \sum_{i=1}^{j} f(x_{i-1}^n) h_n \Big| \le \frac{1}{2}(\bar{x} - a) L_f h_n.$$

Furthermore, if $|f(x)| \le M_f$ for $x \in [a, b]$, then $u : [a, b] \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant M_f and $K_u \le \frac{1}{2} L_f$, where K_u is the constant of uniform differentiability of u.

Additivity:

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

Linearity: If α and β are real numbers then,

$$\int_a^b (\alpha f(x) + \beta g(x))\,dx = \alpha \int_a^b f(x)\,dx + \beta \int_a^b g(x)\,dx.$$

Monotonicity: if $f(x) \ge g(x)$ for $a \le x \le b$, then

$$\int_a^b f(x)\,dx \ge \int_a^b g(x)\,dx.$$

Differentiation and integration are inverse operations:

$$\frac{d}{dx} \int_a^x f(y)\,dy = f(x).$$

Change of variables: Setting $y = g(x)$, we have

$$\int_a^b f(g(x)) g'(x)\,dx = \int_{g(a)}^{g(b)} f(y)\,dy.$$

Integration by parts:

$$\int_a^b u'(x)v(x)\,dx = u(b)v(b) - u(a)v(a) - \int_a^b u(x)v'(x)\,dx.$$

The Mean Value Theorem: If $u(x)$ is uniformly differentiable on $[a, b]$ with Lipschitz continuous derivative $u'(x)$, then there is a (at least one) $\bar{x} \in [a, b]$, such that

$$u(b) - u(a) = (b - a)u'(\bar{x}).$$

Bolzano's theorem: If $f : [a, b] \to \mathbb{R}$ is Lipschitz continuous and $f(a)f(b) < 0$, then there is a real number $\bar{x} \in [a, b]$ such that $f(\bar{x}) = 0$.

Taylor's theorem:

$$u(x) = u(\bar{x}) + u'(\bar{x})(x - \bar{x}) + \cdots + \frac{u^{(n)}(\bar{x})}{n!}(x - \bar{x})^n + \int_{\bar{x}}^x \frac{(x - y)^n}{n!} u^{(n+1)}(y)\,dy.$$

41.10 The logarithm

Definition:

$$\log(x) = \int_1^x \frac{1}{y}\,dy \quad \text{for } x > 0.$$

Properties:

$$\begin{aligned}
\frac{d}{dx}\log(x) &= \frac{1}{x} \quad \text{for } x > 0,\\
\log(ab) &= \log(a) + \log(b) \quad \text{for } a, b > 0,\\
\log(a^r) &= r\log(a), \quad \text{for } r \in \mathbb{R}, a > 0.
\end{aligned}$$

41.11 The exponential

Definition: $\exp(x) = e^x$ is the unique solution of the differential equation $u'(x) = u(x)$ for $x \in \mathbb{R}$ and $u(0) = 1$.

Basic properties:

$$\begin{aligned}
\frac{d}{dx}\exp(x) &= \exp(x),\\
\exp(a + b) &= \exp(a)\exp(b) \quad \text{or } e^{a+b} = e^a e^b,\\
\exp(x) &= \lim_{j\to\infty}(1 + \frac{x}{j})^j.
\end{aligned}$$

The inverse of the exponential is the logarithm:

$$y = \exp(x) \quad \text{if and only if } x = \log(y).$$

The function a^x with $a > 0$:

$$a^x = \exp(x \log(a)), \quad \frac{d}{dx} a^x = \log(a) a^x.$$

41.12 The trigonometric functions

Definition of $\sin(x)$ **and** $\cos(x)$: The initial value problem $u''(x) + u(x) = 0$ for $x > 0$ with $u_0 = 0$ and $u_1 = 1$, has a unique solution, which is denoted by $\sin(x)$. The initial value problem $u''(x) + u(x) = 0$ for $x > 0$ with $u_0 = 1$ and $u_1 = 0$, has a unique solution, which is denoted by $\cos(x)$. The functions $\sin(x)$ and $\cos(x)$ extend to $x < 0$ as solutions of $u''(x) + u(x) = 0$ and are periodic with period 2π, and $\sin(\pi) = 0$, $\cos(\frac{\pi}{2}) = 0$.

Properties:

$$\begin{aligned}
\frac{d}{dx}\sin(x) &= \cos(x),\\
\frac{d}{dx}\cos(x) &= -\sin(x), \cos(-x) = \cos(x),\\
\sin(-x) &= -\sin(x),\\
\cos(\pi - x) &= -\cos(x),\\
\sin(\pi - x) &= \sin(x),\\
\cos(x) &= \sin(\frac{\pi}{2} - x),\\
\sin(x) &= \cos(\frac{\pi}{2} - x),\\
\sin(\frac{\pi}{2} + x) &= \cos(x),\\
\cos(\frac{\pi}{2} + x) &= -\sin(x).
\end{aligned}$$

Definition of $\tan(x)$ **and** $\cot(x)$:

$$\tan(x) = \frac{\sin(x)}{\cos(x)}, \quad \cot(x) = \frac{\cos(x)}{\sin(x)}.$$

Derivatives of $\tan(x)$ **and** $\cot(x)$:

$$\frac{d}{dx}\tan(x) = \frac{1}{\cos^2(x)}, \quad \frac{d}{dx}\cot(x) = -\frac{1}{\sin^2(x)}.$$

Trigonometric formulas:

$$\begin{aligned}
\sin(x+y) &= \sin(x)\cos(y)+\cos(x)\sin(y)\\
\sin(x-y) &= \sin(x)\cos(y)-\cos(x)\sin(y)\\
\cos(x+y) &= \cos(x)\cos(y)-\sin(x)\sin(y)\\
\cos(x-y) &= \cos(x)\cos(y)+\sin(x)\sin(y).\\
\sin(2x) &= 2\sin(x)\cos(x)\\
\cos(2x) &= \cos^2(x)-\sin^2(x) = 2\cos^2(x)-1 = 1-2\sin^2(x)
\end{aligned}$$

$$\begin{aligned}
\cos(x)-\cos(y) &= -2\sin(\frac{x+y}{2})\sin(\frac{x-y}{2}),\\
\tan(x+y) &= \frac{\tan(x)+\tan(y)}{1-\tan(x)\tan(y)},\\
\tan(x-y) &= \frac{\tan(x)-\tan(y)}{1+\tan(x)\tan(y)},\\
\sin(x)+\sin(y) &= 2\sin(\frac{x+y}{2})\cos(\frac{x-y}{2}),\\
\sin(x)-\sin(y) &= 2\cos(\frac{x+y}{2})\sin(\frac{x-y}{2}),\\
\cos(x)+\cos(y) &= 2\cos(\frac{x+y}{2})\cos(\frac{x-y}{2}).
\end{aligned}$$

Inverses of trigonometric functions: The inverse of $f(x)=\sin(x)$ with $D(f)=[\frac{-\pi}{2},\frac{\pi}{2}]$ is $f^{-1}(y)=\arcsin(y)$ with $D(\arcsin)=[-1,1]$. The inverse of $f(x)=\tan(x)$ with $D(f)=(\frac{-\pi}{2},\frac{\pi}{2})$ is $f^{-1}(y)=\arctan(y)$ with $D(\arctan)=\mathbb{R}$. The inverse of $y=f(x)=\cos(x)$ with $D(f)=[0,\pi]$ is $f^{-1}(y)=\arccos(y)$ with $D(\arccos)=[-1,1]$. The inverse of $f(x)=\cot(x)$ with $D(f)=(0,\pi)$ is $f^{-1}(y)=\operatorname{arccot}(y)$ with $D(\operatorname{arccot})=\mathbb{R}$. We have

$$\begin{aligned}
\frac{d}{dy}\arcsin(y) &= \frac{1}{\sqrt{1-y^2}}\\
\frac{d}{dy}\arctan(y) &= \frac{1}{1+y^2}\\
\frac{d}{dy}\arccos(y) &= -\frac{1}{\sqrt{1-y^2}}\\
\frac{d}{dy}\operatorname{arccot}(y) &= -\frac{1}{1+y^2},\\
\arctan(u)+\arctan(v) &= \arctan(\frac{u+v}{1-uv}).
\end{aligned}$$

Definition of sinh(x) and cosh(x):

$$\sinh(x)=\frac{e^x-e^{-x}}{2}\quad\text{and}\quad\cosh(x)=\frac{e^x+e^{-x}}{2}\quad\text{for} x\in\mathbb{R}.$$

Derivatives of of sinh(x) **and cosh**(x):

$$D\sinh(x) = \cosh(x) \quad \text{and} \quad D\cosh(x) = \sinh(x).$$

Inverses of of sinh(x) **and cosh**(x): the inverse of $y = f(x) = \sinh(x)$ with $D(f) = \mathbb{R}$ is $f^{-1}(y) = \operatorname{arcsinh}(y)$ with $D(\operatorname{arcsinh}) = \mathbb{R}$. The inverse of $y = f(x) = \cosh(x)$ with $D([0,\infty))$, is $f^{-1}(y) = \operatorname{arccosh}(y)$ with $D(\operatorname{arccosh}) = [1,\infty)$. We have

$$\frac{d}{dy}\operatorname{arcsinh}(y) = \frac{1}{\sqrt{y^2+1}}, \quad \frac{d}{dy}\operatorname{arccosh}(y) = \frac{1}{\sqrt{y^2-1}}.$$

41.13 List of primitive functions

$$\int_{x_0}^{x} \frac{1}{s-c}\, ds = \log|x-c| - \log|x_0-c|, \quad c \neq 0,$$

$$\int_{x_0}^{x} \frac{s-a}{(s-a)^2+b^2}\, dx = \frac{1}{2}\log((x-a)^2+b^2) - \frac{1}{2}\log((x_0-a)^2+b^2),$$

$$\int_{x_0}^{x} \frac{1}{(s-a)^2+b^2}\, ds = [\frac{1}{b}\arctan(\frac{x-a}{b})] - [\frac{1}{b}\arctan(\frac{x_0-a}{b})], \quad b \neq 0,$$

$$\int_{0}^{x} y\cos(y)\, dy = x\sin(x) + \cos(x) + 1,$$

$$\int_{0}^{x} \sin(\sqrt{y})\, dy = -2\sqrt{x}\cos(\sqrt{x}) + 2\sin(\sqrt{x}),$$

$$\int_{1}^{x} y^2\log(y)\, dy = \frac{x^3}{3}\log(x) - \frac{x^3}{9} + \frac{1}{9}$$

$$\int_{0}^{x} \frac{1}{\sqrt{1-y^2}}\, dy = \arcsin(x) \quad \text{for } x \in (-1,1)$$

$$\int_{0}^{x} \frac{1}{\sqrt{1-y^2}} = \frac{\pi}{2} - \arccos(x)\, dy \quad \text{for } x \in (-1,1)$$

$$\int_{0}^{x} \frac{1}{1+y^2}\, dy = \arctan(x) \quad \text{for } x \in \mathbb{R}$$

$$\int_{0}^{x} \frac{1}{1+y^2}\, dy = \frac{\pi}{2} - \operatorname{arccot}(x) \quad \text{for } x \in \mathbb{R}.$$

41.14 Series

Definition of convergence: A series $\sum_{i=1}^{\infty} a_i$ converges if and only if the sequence $\{s_n\}_{n=1}^{\infty}$ of partial sums $s_n = \sum_{i=1}^{n} a_i$ converges.

Geometric series: $\sum_{i=0}^{\infty} a^i = \frac{1}{1-a}$ if $|a| < 1$.

Basic facts: A positive series $\sum_{i=1}^{\infty} a_i$ converges if and only if the sequence of partial sums is bounded above.

The series $\sum_{i=1}^{\infty} i^{-\alpha}$ converges if and only if $\alpha > 1$.

An absolutely convergent series is convergent.

An alternating series with the property that the modulus of its terms tends monotonically to zero, converges. Example: $\sum_{i=1}^{\infty}(-i)^{-1}$ converges.

41.15 Separable scalar initial value problems

The solution of the separable scalar initial value problem

$$u'(x) = \frac{h(x)}{g(u(x))} \quad \text{for } 0 < x \leq 1,\ u(0) = u_0,$$

where $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$ are given functions, satisfies for $0 \leq x \leq 1$ the algebraic equation

$$G(u(x)) = H(x) + C,$$

where $G(v)$ and $H(x)$ are primitive functions of $g(v)$, and $C = G(u_0) - H(0)$.

42
Analytic geometry in $\mathbb{R}^n$

> I also think that the (mathematical) mine has become too deep and sooner or later it will be necessary to abandon it if new ore-bearing veins shall not be discovered. Physics and Chemistry display now treasures much more brilliant and easily exploitable, thus, apparently, everybody has turned completely in this direction, and possibly posts in Geometry in the Academy of Sciences will some day be like chairs in Arabic Language in universities at present. (Lagrange, 1781)

42.1 Introduction and survey of basic objectives

We now generalize the discussion of analytic geometry to $\mathbb{R}^n$, where n is an arbitrary natural number. Following the pattern set above for $\mathbb{R}^2$ and $\mathbb{R}^3$, we define $\mathbb{R}^n$ to be the set of all possible ordered n-tuples of the form $(x_1, x_2, ..., x_n)$ with $x_i \in \mathbb{R}$ for $i = 1, ..., n$. We refer to $\mathbb{R}^n$ as *n-dimensional Euclidean space*.

We all have a direct concrete experience of $\mathbb{R}^3$ as the three-dimensional space of the real World, and we may think of $\mathbb{R}^2$ as an infinite flat surface, but we don't have a similar experience with for example $\mathbb{R}^4$, except possibly from some science fiction novel with space ships travelling in four-dimensional space-time. Actually, Einstein in his theory of relativity used $\mathbb{R}^4$ as the set of space-time coordinates (x_1, x_2, x_3, x_4) with $x_4 = t$ representing time, but of course had the same difficulty as we all have of "seeing" an object in $\mathbb{R}^4$. In Fig. 42.33, we show a projection into $\mathbb{R}^3$ of a 4-cube in $\mathbb{R}^4$, and we hope the clever reader can "see" the 4-cube.

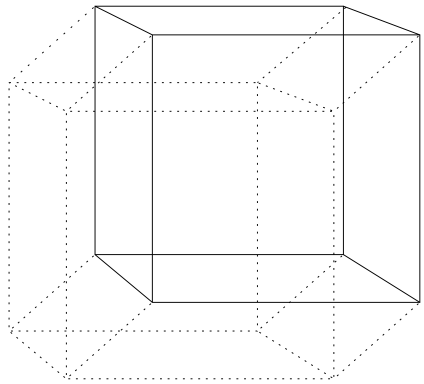


FIGURE 42.1. A cube in $\mathbb{R}^4$

More generally, the need of using $\mathbb{R}^n$ arises as soon as we have n different variables to deal with, which occurs all the time in applications, and $\mathbb{R}^n$ is thus one of the most useful concepts in mathematics. Fortunately, we can work with $\mathbb{R}^n$ purely algebraically without having to draw geometric pictures, that is we can use the tools of analytic geometry in $\mathbb{R}^n$ in pretty much the same way as we have done in $\mathbb{R}^2$ and $\mathbb{R}^3$.

Most of this chapter is one way or the other connected to systems of m linear equations in n unknowns $x_1, ..., x_n$, of the form

$$\sum_{j=1}^{n} a_{ij}x_j = b_i \quad \text{for } i = 1, ..., m, \tag{42.1}$$

that is,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n &= b_2, \\ ...... & \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n &= b_m, \end{aligned} \tag{42.2}$$

where the a_{ij} are given (real) coefficients and $(b_1, ..., b_m) \in \mathbb{R}^m$ is a given right-hand side. We will write this system in matrix form as

$$Ax = b, \tag{42.3}$$

that is

$$\begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ . & . & .. & . \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ . \\ . \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ . \\ . \\ b_m \end{pmatrix}, \tag{42.4}$$

where $A = (a_{ij})$ is a $m \times n$ matrix with rows $(a_{i1}, ..., a_{in})$, $i = 1, ..., m$, and columns $(a_{1j}, ..., a_{mj})$, $j = 1, ..., n$, and we view $x = (x_1, ..., x_n) \in \mathbb{R}^n$ and $b = (b_1, ..., b_m) \in \mathbb{R}^m$ as column vectors. We will also write the system in the form

$$x_1 a_1 + \cdots + x_n a_n = b, \tag{42.5}$$

expressing the given column vector $b \in \mathbb{R}^m$ as a linear combination of the column vectors $a_j = (a_{1j}, a_{2j}, ..., a_{mj})$, $j = 1, 2, ..., n$, with coefficients $(x_1, ..., x_n)$. Notice that we use both (column) vectors in $\mathbb{R}^m$ (such as the columns of the matrix A and the right hand side b) and (column) vectors in $\mathbb{R}^n$ such as the solution vector x.

We shall view $f(x) = Ax$ as a function or transformation $f : \mathbb{R}^n \to \mathbb{R}^m$, and we thus focus on a particular case of our general problem of solving systems of equations of the form $f(x) = b$, where $f : \mathbb{R}^n \to \mathbb{R}^m$ is the *linear transformation* $f(x) = Ax$. We shall denote by $R(A)$ the *range* of $f(x) = Ax$, that is

$$R(A) = \{Ax \in \mathbb{R}^m : x \in \mathbb{R}^n\} = \{\sum_{j=1}^{n} x_j a_j : x_j \in \mathbb{R}\},$$

and by $N(A)$ the *null space* of $f(x) = Ax$ that is

$$N(A) = \{x \in \mathbb{R}^n : Ax = 0\} = \{x \in \mathbb{R}^n : \sum_{j=1}^{n} x_j a_j = 0\}.$$

We are interested in the question of existence and/or uniqueness of solutions $x \in \mathbb{R}^n$ to the problem $Ax = b$ for a given $m \times n$ matrix A and right hand side $b \in \mathbb{R}^m$. Of particular interest is the case $m = n$ with as many equations as unknowns.

Existence of a solution x to $Ax = b$ is of course the same as saying that $b \in R(A)$, which is the same as saying that b is a linear combination of the columns of A. Uniqueness is the same as saying that $N(A) = 0$, because if x and $\hat{x}$ satisfy $Ax = b$ and $A\hat{x} = b$, then by linearity, $A(x - \hat{x}) = 0$, and if $N(A) = 0$ then $x - \hat{x} = 0$ that is $x = \hat{x}$. Further, the non-uniqueness of solutions of $Ax = b$ is described by $N(A)$: If $A\hat{x} = b$ and $Ax = b$, then $x - \hat{x} \in N(A)$.

We may thus formulate the following prime objectives of our study of the linear transformation $f(x) = Ax$ given by the matrix A:

- Determine $R(A)$.
- Determine $N(A)$.
- Solve $Ax = b$ for given b.

We state here the following partial answer given by the *Fundamental Theorem of Linear Algebra*, which we will prove in a couple of different ways below: Let $m = n$ and suppose that $N(A) = 0$. Then $Ax = b$ has a unique solution for any $b \in \mathbb{R}^m$, that is, $R(A) = \mathbb{R}^m$. In other words, if $m = n$, then uniqueness implies existence.

In our study we will be led to concepts such as: linear combination, linear span, linear space, vector space, subspace, linear independence, basis, determinant, linear transformation and projection, which we have already met in the chapters on analytic geometry in $\mathbb{R}^2$ and $\mathbb{R}^3$ above.

This chapter focusses mostly on theoretical issues while the computational methods such as Gaussian elimination and iterative methods are considered in more detail in Chapter *Solving systems of linear equations* below.

42.2 Body/Soul and Artificial Intelligence

Before plunging into the geometry of $\mathbb{R}^n$, we take a brake and return to the story of Body and Soul which continues into our time with new questions: Is it possible to create computer programs for Artificial Intelligence AI, that is, can we give the computer some more or less advanced capability of acting like an intelligent organism with some ability of "thinking"? It appears that this question does not yet have a clear positive answer, despite many dreams in that direction during the development of the computer. In seeking an answer, Spencer's principle of adaptivity of course plays an important role: an intelligent system must be able to adapt to changes in its environment. Further, the presence of a goal or final cause according to Leibniz, seems to be an important feature of intelligence, to judge if an action of a system is stupid or not. Below we will design adaptive IVP-solvers, which are computer programs for solving systems of differential equations, with features of adaptive feed-back from the computational process towards the goal of error control. These IVP-solvers thus show exhibit some kind of rudiments of intelligence, and at any rate are infinitely much more "clever" than traditional non-adaptive IVP-solvers with no feed-back.

42.3 $\mathbb{R}^n$ as a vector space. Addition and scalar multiplication.

We view $\mathbb{R}^n$ as a *vector space* consisting of *vectors* which are ordered n-tuples, $x = (x_1, ..., x_n)$ with components $x_i \in \mathbb{R}$, $i = 1, ..., n$. We write $x = (x_1, ..., x_n)$ for short, and refer to $x \in \mathbb{R}^n$ as a vector with component x_i in position i.

We may *add* two vectors $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ in $\mathbb{R}^n$ by componentwise addition to get a new vector $x + y$ in $\mathbb{R}^n$ defined by

$$x + y = (x_1 + y_1, x_2 + y_2, ..., x_n + y_n). \tag{42.6}$$

Further, we may multiply a vector $x = (x_1, ..., x_n)$ by a real number λ by componentwise multiplication with λ, to get a new vector λx in $\mathbb{R}^n$ defined by

$$\lambda x = (\lambda x_1, ..., \lambda x_n). \tag{42.7}$$

The operations of adding two vectors in $\mathbb{R}^n$ and multiplying a vector in $\mathbb{R}^n$ with a real number, of course directly generalize the corresponding operations from the cases $n = 2$ and $n = 3$ considered above. The generalization helps us to deal with $\mathbb{R}^n$ using concepts and tools which we have found useful in $\mathbb{R}^2$ and $\mathbb{R}^3$.

We may thus add vectors in $\mathbb{R}^n$ and multiply them by real numbers (scalars), the usual commutative and distributive rules hold for these operations, and $\mathbb{R}^n$ is thus a vector space. We say that $(0, 0, .., 0)$ is the *zero vector* in $\mathbb{R}^n$ and write $0 = (0, 0, .., 0)$.

Linear algebra concerns vectors in vector spaces, also refereed to as *linear spaces*, and linear functions of vectors, that is *linear transformations* of vectors. As we just saw, $\mathbb{R}^n$ is a vector space, but there are also many other types of vector spaces, where the vectors have a different nature. In particular, we will below meet vector spaces consisting of vectors which are functions. In this chapter we focus on $\mathbb{R}^n$, the most basic of all vector spaces. We know that linear transformations in $\mathbb{R}^2$ and $\mathbb{R}^3$ lead to 2×2 and 3×3 matrices, and we shall now generalize to linear transformations from $\mathbb{R}^n$ into $\mathbb{R}^m$ which can be represented by $m \times n$ *matrices*.

We give in this chapter a condensed (and dry) presentation of some basic facts of linear algebra in $\mathbb{R}^n$. Many applications of the theoretical results presented will appear in the rest of the book.

42.4 Scalar product, orthogonality

We define the *scalar product* $x \cdot y = (x, y)$ of two vectors x and y in $\mathbb{R}^n$, by

$$x \cdot y = (x, y) = \sum_{i=1}^{n} x_i y_i. \tag{42.8}$$

This generalizes the scalar product in $\mathbb{R}^2$ and $\mathbb{R}^3$. Note that here we introduce a new notation for the scalar product of two vectors x and y, namely (x, y), as an alternative to the "dot product" $x \cdot y$ used in $\mathbb{R}^2$ and $\mathbb{R}^3$. We should be ready to use both notations.

The scalar product is *bilinear* in the sense that $(x + y, z) = (x, z) + (y, z)$, $(\lambda x, z) = \lambda(x, z)$, $(x, y + z) = (x, y) + (x, z)$ and $(x, \lambda y) = \lambda(x, y)$, and *symmetric* in the sense that $(x, y) = (y, x)$, for all vectors $x, y, z \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$.

We say that two vectors x and y in $\mathbb{R}^n$ are *orthogonal* if $(x, y) = 0$. We define

$$|x| = (\sum_{i=1}^{n} x_i^2)^{1/2} = (x, x)^{1/2} \tag{42.9}$$

to be the Euclidean *length* or *norm* of the vector x. Note that this definition of *length* is a direct generalization of the natural length $|x|$ of a vector x in $\mathbb{R}^n$, $n = 1, 2, 3$.

EXAMPLE 42.1. Let $x = (2, -4, 5, 1, 3)$ and $y = (1, 4, 6, -1, 2)$ be two vectors in $\mathbb{R}^5$. We compute $(x, y) = 2 \times 1 + (-4) \times 4 + 5 \times 6 + 1 \times (-1) + 3 \times 2 = 21$.

42.5 Cauchy's inequality

Cauchy's inequality states that for $x, y \in \mathbb{R}^n$,

$$|(x, y)| \leq |x|\,|y|.$$

In words: the absolute value of the scalar product of two vectors is bounded by the product of the norms of the vectors. We prove Cauchy's inequality by noting that for all $s \in \mathbb{R}$,

$$0 \leq |x + sy|^2 = (x + sy, x + sy) = |x|^2 + 2s(x, y) + s^2|y|^2,$$

and then assuming that $y \neq 0$, choosing $s = -(x, y)/|y|^2$ (which minimizes the right-hand side), to get

$$0 \leq |x|^2 - 2\frac{(x, y)^2}{|y|^2} + \frac{(x, y)^2}{|y|^2} = |x|^2 - \frac{(x, y)^2}{|y|^2},$$

which proves the desired result.

We recall that for $n = 2, 3$,

$$(x, y) = x \cdot y = \cos(\theta)|x||y|,$$

where θ is the angle between x and y, from which of course Cauchy's inequality follows directly using the fact that $|\cos(\theta)| \leq 1$.

We define the *angle* $\theta \in [0, 2\pi)$ between two non-zero vectors x and y in $\mathbb{R}^n$ by

$$\cos(\theta) = \frac{(x, y)}{|x||y|}, \tag{42.10}$$

which generalizes the corresponding notion for $n = 2, 3$.

EXAMPLE 42.2. The angle between the vectors $x = (1, 2, 3, 4)$ and $y = (4, 3, 2, 1)$ in $\mathbb{R}^4$ is equal to $\arccos \frac{2}{3} \approx 0.8411 \approx 48°$ since $(x, y) = 20$ and $|x| = |y| = \sqrt{30}$.

42.6 Vector space of linear combinations of a set of vectors

We know that two non-parallel vectors a_1 and a_2 in $\mathbb{R}^3$ define a plane in $\mathbb{R}^3$ through the origin consisting of all the linear combinations $\lambda_1 a_1 + \lambda_2 a_2$ with coefficients λ_1 and λ_2 in $\mathbb{R}$. The normal to the plane is given by $a_1 \times a_2$. A plane through the origin is an example of *subspace* of $\mathbb{R}^3$, which is a subset of $\mathbb{R}$ with the property that vector addition and scalar multiplication does not lead outside the set. So, a subset S of $\mathbb{R}^3$ is a subspace if the sum of any two vectors in S belongs to S and scalar multiplication of a vector in S gives a vector in S. Clearly, a plane through the origin is a subspace of $\mathbb{R}^3$. Similarly, a line through the origin defined as the scalar multiples $\lambda_1 a_1$ with coefficients $\lambda_1 \in \mathbb{R}$ and a_1 a given vector in $\mathbb{R}^3$, is a subspace of $\mathbb{R}^3$. The subspaces of $\mathbb{R}^3$ consist of lines and planes through the origin. Notice that a plane or line in $\mathbb{R}^3$ not passing through the origin, is not a subspace.

More generally, we use the concept of a *vector space* to denote a set of vectors for which the operations of vector addition and scalar multiplication does not lead outside the set. Of course, $\mathbb{R}^3$ is a vector space. A subspace of $\mathbb{R}^3$ is a vector space. A plane or line in $\mathbb{R}^3$ through the origin is a vector space. The concept of vector space is fundamental in mathematics and we will meet this term many times below.

We will now generalize to $\mathbb{R}^m$ with $m > 3$ and we will then meet new examples of vector spaces and subspaces of vector spaces. Let a_1, a_2,...,a_n, be n non-zero vectors in $\mathbb{R}^m$. A vector b in $\mathbb{R}^m$ of the form

$$b = \lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_n a_n, \tag{42.11}$$

where the $\lambda_i \in \mathbb{R}$, is said to be a *linear combination* of the set of vectors $\{a_1, ..., a_n\}$ with *coefficients* $\lambda_1, ..., \lambda_n$. If

$$c = \mu_1 a_1 + \mu_2 a_2 + \cdots + \mu_n a_n, \tag{42.12}$$

is another linear combination of $\{a_1, ..., a_n\}$ with coefficients $\mu_j \in \mathbb{R}$, then the vector

$$b + c = (\lambda_1 + \mu_1)a_1 + (\lambda_2 + \mu_2)a_2 + \cdots + (\lambda_+ \mu_n)a_n, \tag{42.13}$$

is again a linear combination of $\{a_1, ..., a_m\}$ now with coefficients $\lambda_j + \mu_j$. Further, for any $\alpha \in \mathbb{R}$ the vector

$$\alpha b = \alpha\lambda_1 a_1 + \alpha\lambda_2 a_2 + \cdots + \alpha\lambda_m a_m \tag{42.14}$$

is also a linear combination of $\{a_1, ..., a_m\}$ with coefficients $\alpha\lambda_j$. This means that if we let $S(a_1, ...a_n)$ denote the set of all linear combinations

$$\lambda_1 a_1 + \lambda_2 a_2 + \cdots + \lambda_n a_n, \tag{42.15}$$

of $\{a_1, ..., a_n\}$, where the coefficients $\lambda_j \in \mathbb{R}$, then $S(a_1, ...a_n)$ is indeed a vector space, since vector addition and multiplication by scalars do not lead outside the set. The sum of two linear combinations of $\{a_1, ..., a_n\}$ is also a linear combination of $\{a_1, ..., a_n\}$, and a linear combination of $\{a_1, ..., a_n\}$ multiplied by a real number is also a linear combination of $\{a_1, ..., a_n\}$.

We refer to the vector space $S(a_1, ...a_n)$ of all linear combinations of the form (42.15) of the vectors $\{a_1, ..., a_n\}$ in $\mathbb{R}^m$ as the *subspace of* $\mathbb{R}^m$ *spanned by the vectors* $\{a_1, ..., a_n\}$, or simply just the *span of* $\{a_1, ..., a_n\}$, which we may describe as:

$$S(a_1, ...a_n) = \{\sum_{i=1}^{n} \lambda_j a_j : \ \lambda_j \in \mathbb{R}, j = 1, ..., n\}.$$

If $m = 2$ and $n = 1$, then the subspace $S(a_1)$ is a line in $\mathbb{R}^2$ through the origin with direction a_1. If $m = 3$ and $n = 2$, then $S(a_1, a_2)$ corresponds to the plane in $\mathbb{R}^3$ through the origin spanned by a_1 and a_2 (assuming a_1 and a_2 are non-parallel), that is, the plane through the origin with normal given by $a_1 \times a_2$.

Note that for any $\mu \in \mathbb{R}$, we have

$$S(a_1, a_2, ..., a_n) = S(a_1, a_2 - \mu a_1, a_3, ..., a_n), \tag{42.16}$$

since we can replace each occurence of a_2 by the linear combination $(a_2 - \mu a_1) + \mu a_1$ of $a_2 - \mu a_1$ and a_1. More generally, we can add any multiple of one vector to one of the other vectors without changing the span of the vectors! Of course we may also replace any vector a_j with a μa_j where μ is a non-zero real number without changing the span. We shall return to these operations below.

42.7 The standard basis

The set of vectors in $\mathbb{R}^n$:

$$\{(1, 0, 0, \cdots, 0, 0), (0, 1, 0, \cdots, 0, 0), \cdots, (0, 0, 0, \cdots, 0, 1)\},$$

commonly denoted by $\{e_1, ..., e_n\}$, where $e_i = (0, 0, .., 0, 1, 0, ..., 0)$ with a single coefficient 1 at position i, is called the *standard basis* for $\mathbb{R}^n$. Any vector $x = (x_1, ..., x_n) \in \mathbb{R}^n$ can be written as a linear combination of the basis vectors $\{e_1, ..., e_n\}$:

$$x = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n, \tag{42.17}$$

with the coefficients x_j of x appearing as coefficients of the basis vectors e_j. We note that $(e_j, e_k) = e_j \cdot e_k = 0$ for $j \neq k$, that is the standard basis vectors are *pairwise orthogonal*, and of length one since $(e_j, e_j) = |e_j|^2 = 1$.

We may thus express the cofficients x_i of a given vector $x = (x_1, ..., x_n)$ with respect to the standard basis $\{e_1, ..., e_n\}$ as follows:

$$x_i = (e_i, x) = e_i \cdot x. \tag{42.18}$$

42.8 Linear independence

We recall that to specify a plane in $\mathbb{R}^3$ as the set of linear combinations of two given vectors a_1 and a_2, we assume that a_1 and a_2 are non-parallel. The generalization of this condition to a set $\{a_1, ..., a_m\}$ of m vectors in $\mathbb{R}^n$, is referred to as *linear independence*, which we now proceed to define. Eventually, this will lead us to the concept of *basis* of a vector space, which is one of the most basic(!) concepts of linear algebra.

A set $\{a_1, ..., a_n\}$ of vectors in $\mathbb{R}^m$ is said to be *linearly independent* if none of the vectors a_i can be expressed as a linear combination of the others. Conversely, if some of the vectors a_i can be expressed as a linear combination of the others, for example if

$$a_1 = \lambda_2 a_2 + ... + \lambda_n a_n \tag{42.19}$$

for some numbers λ_2,...,λ_n, we say that the set $\{a_1, a_2, ..., a_n\}$ is *linearly dependent*. As a test of linear independence of $\{a_1, a_2, ..., a_n\}$, we can use: if

$$\lambda_1 a_1 + \lambda_2 a_2 + ... + \lambda_n a_n = 0 \tag{42.20}$$

implies that $\lambda_1 = \lambda_2 = ... = \lambda_n = 0$, then $\{a_1, a_2, ..., a_n\}$ is linearly independent. This is because if (42.20) holds with some of the λ_j different from 0, for example, $\lambda_1 \neq 0$, then we could divide by λ_1 and express a_1 as a linear combination of $\{a_2, ..., a_m\}$:

$$a_1 = -\frac{\lambda_2}{\lambda_1} a_2 + ... + -\frac{\lambda_n}{\lambda_1} a_n. \tag{42.21}$$

The standard basis $\{e_1, ..., e_n\}$ is (of course) a linearly independent set, since if

$$\lambda_1 e_1 + ... + \lambda_n e_n = 0,$$

then $\lambda_i = 0$ for $i = 1, ..., n$, because $0 = (0, 0, ..., 0) = \lambda_1 e_1 + ... + \lambda_n e_n = (\lambda_1, ..., \lambda_n)$.

42.9 Elimination of linearly dependent elements gives a basis of a subspace

Consider the subspace $S(a_1, ..., a_n)$ spanned by the set of vectors $\{a_1, a_2, ..., a_n\}$. If the set $\{a_1, a_2, ..., a_n\}$ is linearly dependent, say that a_n can be expressed as a linear combination of $\{a_1, ..., a_{n-1}\}$, then $S(a_1, ..., a_n)$ is in fact spanned by $\{a_1, ..., a_{n-1}\}$ and thus $S(a_1, ..., a_n) = S(a_1, ..., a_{n-1})$. This follows simply by replacing all occurences of a_n by its linear combination of $\{a_1, ..., a_{n-1}\}$. Continuing this way, eliminating linearly dependent vectors, we may express $S(a_1, ..., a_n)$ as the span of $\{a_1, a_2, ..., a_k\}$ (with a suitable renumeration), that is, $S(a_1, ..., a_n) = S(a_1, a_2, ..., a_k\}$, where $k \leq n$, and the set $\{a_1, a_2, ..., a_k\}$ is linearly independent. This means that $\{a_1, a_2, ..., a_k\}$ is a *basis* for the vector space $S = S(a_1, ..., a_n)$ in the sense that the following two conditions are fulfilled:

- any vector in S can be expressed as a linear combination of $\{a_1, a_2, ..., a_k\}$,
- the set $\{a_1, a_2, ..., a_k\}$ is linearly independent.

Note that by the linear independence the coefficients in the linear combination are uniquely determined: if two linear combinations $\sum_{j=1}^{k} \lambda_j a_j$ and $\sum_{j=1}^{k} \mu_j a_j$ are equal, then $\lambda_j = \mu_j$ for $j = 1, .., k$.

Each vector $b \in S$ can thus be expresses as a unique linear combination of the basis vectors $\{a_1, a_2, ..., a_k\}$:

$$b = \sum_{j=1}^{k} \lambda_j a_j,$$

and we refer to $(\lambda_1, ..., \lambda_k)$ as the *coefficients* of b with respect to the basis $\{a_1, a_2, ..., a_k\}$.

The *dimension* of a vector space S is equal to the number of basis vectors in a basis for S. We prove below that the dimension is uniquely defined so that two sets of basis vectors always have the same number of elements.

EXAMPLE 42.3. Consider the three vectors $a_1 = (1, 2, 3, 4)$, $a_2 = (1, 1, 1, 1)$, and $a_3 = (3, 3, 5, 6)$ in $\mathbb{R}^4$. We see that $a_3 = a_1 + 2a_2$, and thus the set $\{a_1, a_2, a_3\}$ is linearly dependent. The span of $\{a_1, a_2, a_3\}$ thus equals the span of $\{a_1, a_2\}$, since each occurrence of a_3 can be replaced by $a_1 + 2a_2$. The vector a_3 is thus redundant, since it can be replaced by a linear combination of a_1 and a_2. Evidently, $\{a_1, a_2\}$ is linearly independent, since a_1 and a_2 are non-parallel. Thus, $\{a_1, a_2\}$ is a linearly independent set spanning the same subset as $\{a_1, a_2, a_3\}$. We can also express a_2 in terms of a_1 and a_3, or a_1 in terms of a_2 and a_3, and thus any set of two vectors $\{a_1, a_2\}$, $\{a_1, a_3\}$ or $\{a_2, a_3\}$, can serve as a basis for the subspace spanned by $\{a_1, a_2, a_3\}$.

42.10 Constructing a basis for $S(a_1, ..., a_n)$ by reduction to column echelon form

We now present a constructive process for determining a basis for the vector space $S(a_1, ..., a_n)$ spanned by the set of vectors $\{a_1, a_2, ..., a_n\}$, where

$a_j = (a_{1j}, ..., a_{mj}) \in \mathbb{R}^m$ for $j = 1, ..., n$ which we view as column vectors. We refer to this process as *reduction to column echelon form.* It is of fundamental importance and we shall return to it below in several different contexts. Assume then first that $a_{11} = 1$ and choose $\mu \in \mathbb{R}$ so that $\mu a_{11} = a_{12}$, and note that $S(a_1, ..., a_n) = S(a_1, a_2 - \mu a_1, a_3, ..., a_n)$, where now the first component of $a_2 - \mu a_1$ is zero. We here used the fact that we can add one vector multiplied by a scalar to another vector without changing the span of the vectors. Continuing in the same way we obtain $S(a_1, ..., a_n) = S(a_1, \hat{a}_2, \hat{a}_3, ...\hat{a}_n)$ where $\hat{a}_{1j} = 0$ for $j > 1$. In matrix form with the $a_j \in \mathbb{R}^m$ as column vectors, we may express this as follows:

$$S\begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ a_{21} & a_{22} & .. & a_{2n} \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix} = S\begin{pmatrix} 1 & 0 & .. & 0 \\ a_{21} & \hat{a}_{22} & .. & \hat{a}_{22} \\ . & . & .. & . \\ a_{m1} & \hat{a}_{m2} & .. & \hat{a}_{mn} \end{pmatrix}$$

We can now repeat the process by cutting out the first row and first column and reduce to a set of $n-1$ vectors in $\mathbb{R}^{m-1}$. Before doing this we take care of the case $a_{11} \neq 1$. If $a_{11} \neq 0$, then we transform to the case $a_{11} = 1$ by replacing a_1 by μa_1 with $\mu = 1/a_{11}$, noting that we can multiply any column with a non-zero real number without changing the span. By renumbering the vectors we may then assume that either $a_{11} \neq 0$, which thus led to the above construction, or $a_{1j} = 0$ for $j = 1, ..., n$, in which case we seek to compute a basis for

$$S\begin{pmatrix} 0 & 0 & .. & 0 \\ a_{21} & a_{22} & .. & a_{22} \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix}$$

with only zeros in the first row. We may then effectively cut out the first row and reduce to a set of n vectors in $\mathbb{R}^{m-1}$.

Repeating now the indicated process, we obtain with $k \leq \min(n, m)$,

$$S\begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ . & . & .. & . \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix} = S\begin{pmatrix} 1 & 0 & 0 & .. & 0 & .. & 0 \\ \hat{a}_{21} & 1 & 0 & .. & 0 & .. & 0 \\ . & . & . & 0 & 0 & .. & 0 \\ . & . & .. & 1 & 0 & .. & 0 \\ . & . & .. & . & . & .. & . \\ \hat{a}_{m1} & \hat{a}_{m2} & .. & \hat{a}_{mk} & 0 & .. & 0 \end{pmatrix},$$

where we refer to the matrix on the right as the *column echelon form* of the matrix to the left, and k is the number of non-zero columns. We see that each non-zero column $\hat{a}_j$, $j = 1, ..., k$, in the echelon form has a coefficient equal to 1 and that all matrix elements to the right and above that coefficient is equal to zero. Further, the ones appear in a staircase form descending to the right on or below the diagonal. The set of non-zero columns $\{\hat{a}_1, ..., \hat{a}_k\}$ is linearly independent, because if

$$\sum_{j=1}^{k} \hat{x}_j \hat{a}_j = 0,$$

then we get successively $\hat{x}_1 = 0$, $\hat{x}_2 = 0$,..., $\hat{x}_k = 0$, and thus $\{\hat{a}_1, ..., \hat{a}_k\}$ forms a basis for $S(a_1, ..., a_n)$. The dimension of $S(a_1, ..., a_n)$ is equal to k. If zero columns appear in the echelon form, then the original set $\{a_1, ..., a_n\}$ is linearly dependent.

We note that, because of the construction, zero columns must appear if $n > m$, and we thus understand that a set of n vectors in $\mathbb{R}^m$ is linearly dependent if $n > m$. We may also understand that if $n < m$, then the set $\{a_1, ..., a_n\}$ cannot span $\mathbb{R}^m$, because if $k < m$, then there are vectors $b \in \mathbb{R}^m$ which cannot be expressed as linear combinations of $\{\hat{a}_1, ..., \hat{a}_k\}$ as we now show: if

$$b = \sum_{j=1}^{k} \hat{x}_j \hat{a}_j,$$

then the coefficients $\hat{x}_1, ..., \hat{x}_k$ are determined by the coefficients $b_1, ..., b_k$, of b occurring in the rows with the coefficient 1. For example, in the case the 1s appear on the diagonal, we first compute $\hat{x}_1 = b_1$, then $\hat{x}_2 = b_1 - \hat{a}_{21}\hat{x}_1$ etc, and thus the remaining coefficients $b_{k+1}, ..., b_m$ of b cannot be arbitrary.

42.11 Determining $R(A)$ by reduction to column echelon form

By reduction to column echelon form we can construct a basis for $R(A)$ for a given $m \times n$ matrix A with column vectors $a_1, ..., a_n$ because

$$Ax = \sum_{j=1}^{n} x_j a_j$$

and thus $R(A) = S(a_1, ..., a_n)$ expressing that the range $R(A) = \{Ax : x \in \mathbb{R}^n\}$ is equal to the vector space $S(a_1, ..., a_n)$ of all linear combinations of the set of column vectors $\{a_1, ..., a_n\}$. Setting now

$$A = \begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ . & . & .. & . \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix}, \qquad \hat{A} = \begin{pmatrix} 1 & 0 & 0 & .. & 0 & .. & 0 \\ \hat{a}_{21} & 1 & 0 & .. & 0 & .. & 0 \\ . & . & . & 0 & 0 & .. & 0 \\ . & . & .. & 1 & 0 & .. & 0 \\ . & . & .. & . & . & .. & . \\ \hat{a}_{m1} & \hat{a}_{m2} & .. & \hat{a}_{mk} & 0 & .. & 0 \end{pmatrix}$$

with $\hat{A}$ obtained from A by reduction to column echelon form, we have

$$R(A) = R(\hat{A}) = S(\hat{a}_1, ..., \hat{a}_k),$$

and thus $\{\hat{a}_1, ..., \hat{a}_k\}$ forms a basis for $R(A)$. In particular we can easily check if a given vector $b \in \mathbb{R}^m$ belongs to $R(A)$, by using the echelon form. By reduction to column echelon form we can thus give an answer to the basic problem of determining $R(A)$ for a given matrix A. Not bad. For example, in the case $m = n$ we have that $R(A) = \mathbb{R}^m$ if and only if $k = n = m$, in which case the echelon form $\hat{A}$ has 1s all along the diagonal.

We give an example showing the sequence of matrices appearing in reduction to column echelon form:

EXAMPLE 42.4. We have

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 7 \\ 1 & 3 & 4 & 5 & 8 \\ 1 & 4 & 5 & 6 & 9 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 2 & 3 & 6 \\ 1 & 2 & 3 & 4 & 7 \\ 1 & 3 & 4 & 5 & 8 \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & -1 & -2 & -5 \\ 1 & 3 & -2 & -4 & -10 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & -2 & -5 \\ 1 & 3 & 2 & -4 & -10 \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 1 & 3 & 2 & 0 & 0 \end{pmatrix} = \hat{A}.$$

We conclude that $R(A)$ is spanned by the 3 non-zero columns of $\hat{A}$ and thus in particular that the dimension of $R(A)$ is equal to 3. In this example, A is a 4×5 matrix and $R(A)$ does not span $\mathbb{R}^4$. Solving the system

$$\hat{A}\hat{x} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 1 & 3 & 2 & 0 & 0 \end{pmatrix} \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \\ \hat{x}_4 \\ \hat{x}_5 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix}$$

we compute uniquely $\hat{x}_1$, $\hat{x}_2$ and $\hat{x}_3$ from the first three equations, and to have the fourth equation satisfied, we must have $b_4 = \hat{x}_1 + 3\hat{x}_2 + 2\hat{x}_3$ and thus b_4 can not be chosen freely.

42.12 Determining $N(A)$ by reduction to row echelon form

We take the chance to solve the other basic problem of determining $N(A)$ by *reduction to row echelon form*, which is analogous to reduction to column echelon form working now with the rows instead of the columns. We thus consider a $m \times n$ matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ . & . & .. & . \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix},$$

and perform the operations of (i) multiplying one row with a real number and (ii) multiplying one row with a real number and subtracting it from another row. We then obtain the *row echelon form* of A (possibly by reordering rows):

$$\hat{A} = \begin{pmatrix} 1 & \hat{a}_{12} & . & . & .. & . & \hat{a}_{1n} \\ 0 & 1 & . & .. & . & .. & \hat{a}_{2n} \\ . & . & . & . & . & .. & . \\ 0 & 0 & .. & 1 & . & .. & \hat{a}_{kn} \\ 0 & 0 & .. & 0 & 0 & .. & 0 \\ . & . & .. & . & . & .. & . \\ 0 & 0 & .. & 0 & 0 & .. & 0 \end{pmatrix}$$

Each non-zero row of the row echelon matrix $\hat{A}$ has one element equal to 1 and all elements to the left and below are equal to zero, and the 1s appear in a staircase form on or to the right of the diagonal from the upper left corner.

We notice that the row operations do not change the null space $N(A) = \{x : Ax = 0\}$, because we may perform the row operations in the system of equations $Ax = 0$, that is

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n &= 0, \\ a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n &= 0, \\ ...... \\ a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n &= 0, \end{aligned} \tag{42.22}$$

to reduce it to the echelon form system $\hat{A}x = 0$ without changing the vector $x = (x_1, ..., x_n)$. We conclude that

$$N(A) = N(\hat{A})$$

and we may thus determine $N(A)$ by using that we can directly determine $N(\hat{A})$ from the echelon form of A. It is easy to see that the dimension of

$N(A) = N(\hat{A})$ is equal to $n - k$, as illustrated in the following example. In the case $n = m$, we have that $N(A) = 0$ if and only if $k = m = n$ in which all diagonal elements of $\hat{A}$ are equal to 1.

We give an example showing the sequence of matrices appearing in reduction to row echelon form:

Example 42.5. We have

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 7 \\ 1 & 3 & 4 & 5 & 8 \\ 1 & 4 & 5 & 6 & 9 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 6 \\ 0 & 2 & 3 & 4 & 7 \\ 0 & 3 & 4 & 5 & 8 \end{pmatrix} \rightarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 6 \\ 0 & 0 & -1 & -2 & -5 \\ 0 & 0 & -2 & -4 & -10 \end{pmatrix}$$

$$\rightarrow \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 6 \\ 0 & 0 & 1 & 2 & 5 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} = \hat{A}.$$

We now determine $N(A)$ by determining $N(\hat{A}) = N(A)$ by seeking the solutions $x = (x_1, ..., x_5)$ of the system $\hat{A}x = 0$, that is

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 6 \\ 0 & 0 & 1 & 2 & 5 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

We see that we can freely choose x_4 and x_5 and then solve for x_3, x_2 and x_1 to get the solution in the form

$$x = \lambda_1 \begin{pmatrix} 0 \\ 1 \\ -2 \\ 1 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 0 \\ 4 \\ -5 \\ 0 \\ 1 \end{pmatrix}$$

where λ_1 and λ_2 are any real numbers. We have now computed a basis for $N(A)$ and we see in particular that the dimension of $N(A)$ is equal to 2. We recall that the dimension of $R(A)$ is equal to 3 and we note that the sum of the dimensions of $R(A)$ and $N(A)$ happens to be equal to 5 that is the number of columns of A. This is a general fact which we prove in the Fundamental Theorem below.

42.13 Gaussian elimination

Gaussian elimination to compute solutions to the system

$$\begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ . & . & .. & . \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ . \\ . \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ . \\ . \\ b_m \end{pmatrix}$$

closely couples to reduction to row echelon form. Performing row operations we may reduce to a system of the form

$$\hat{A} = \begin{pmatrix} 1 & \hat{a}_{12} & . & . & . & .. & . & \hat{a}_{1n} \\ 0 & 1 & . & .. & . & .. & \hat{a}_{2n} \\ . & . & . & . & . & .. & . \\ 0 & 0 & .. & 1 & . & .. & \hat{a}_{kn} \\ 0 & 0 & .. & 0 & 0 & .. & 0 \\ . & . & .. & . & . & .. & . \\ 0 & 0 & .. & 0 & 0 & .. & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ . \\ . \\ x_n \end{pmatrix} = \begin{pmatrix} \hat{b}_1 \\ . \\ . \\ \hat{b}_m \end{pmatrix}$$

with the same solution vector x. We may assume, by possibly by renumbering the components of x, that the 1s appear on the diagonal. We see that solvability is equivalent to having $\hat{b}_j = 0$ for $j = k+1, .., m$, and the non-uniqueness is expressed by $N(A)$ as explained above. In the case $m = n$ we have that $N(A) = 0$ if and only if $k = m = n$ in which case all diagonal elements of $\hat{A}$ are equal to 1, and the system $\hat{A}x = \hat{b}$ is uniquely solvable for all $\hat{b} \in \mathbb{R}^m$, and thus $Ax = b$ is uniquely solvable for all $b \in \mathbb{R}^m$. We conclude that if $m = n$, then uniqueness implies existence. We may thus say that by Gaussian elimination or reduction to row echelon form, we may solve our basic problems of existence and uniqueness of solutions to the system $Ax = b$. We shall add more information on these problems in the Fundamental Theorem of Linear Algebra below. For more information on Gaussian elimination, we refer to Chapter *Solving linear algebraic systems* below.

42.14 A basis for $\mathbb{R}^n$ contains n vectors

Let us now prove that if $\{a_1, ..., a_m\}$ is a basis for $\mathbb{R}^n$, then $m = n$, that is any basis for $\mathbb{R}^n$ has exactly n elements, no more no less. We already deduced this fact from the column echelon form above, but we here give a "coordinate-free" proof which applies to more general situations.

We recall that a set $\{a_1, ..., a_m\}$ of vectors in $\mathbb{R}^n$ is a *basis* for $\mathbb{R}^n$ if the following two conditions are fullfilled:

- $\{a_1, ..., a_m\}$ is linearly independent,

- any vector $x \in \mathbb{R}^n$ can be expressed as a linear combination $x = \sum_{j=1}^{m} \lambda_j a_j$ of $\{a_1, ..., a_m\}$ with coefficients λ_j.

Of course, $\{e_1, ..., e_n\}$ is a basis for $\mathbb{R}^n$ in this sense.

To prove that $m = n$, we consider the set $\{e_1, a_1, a_2, ..., a_m\}$. Since $\{a_1, ..., a_m\}$ is a basis for $\mathbb{R}^n$, that is spans $\mathbb{R}^n$, the vector e_1 can be expressed as a linear combination of $\{a_1, ..., a_m\}$:

$$e_1 = \sum_{j=1}^{m} \lambda_j a_j,$$

with some $\lambda_j \neq 0$. Suppose $\lambda_1 \neq 0$. Then, dividing by λ_1 expresses a_1 as a linear combination of $\{e_1, a_2, ..., a_m\}$. This means that $\{e_1, a_2, ..., a_m\}$ spans $\mathbb{R}^n$. Consider now the set $\{e_1, e_2, a_2, ..., a_m\}$. The vector e_2 can be expressed as a linear combination of $\{e_1, a_2, ..., a_m\}$ and some of the coefficients of the a_j must be non-zero, since $\{e_1, e_2\}$ are linearly independent. Supposing the coefficient of a_2 is non-zero, we can eliminate a_2 and thus the set $\{e_1, e_2, a_3, ..., a_m\}$ now spans $\mathbb{R}^n$. Continuing this way we get the set $\{e_1, e_2, ..., e_n, a_{n+1}, ..., a_m\}$ if $m > n$ and the set $\{e_1, e_2, ..., e_n\}$ if $m = n$, which both span $\mathbb{R}^n$. We conclude that $m \geq n$, since if e.g. $m = n - 1$, we would end up with the set $\{e_1, e_2, ..., e_{n-1}\}$ which does not span $\mathbb{R}^n$ contrary to the assumption.

Repeating this argument with the roles of the basis $\{e_1, e_2, ..., e_n\}$ and $\{a_1, a_2, ..., a_m\}$ interchanged, we get the reverse inequality $n \geq m$ and thus $n = m$. Of course, intuitively, there are n independent directions in $\mathbb{R}^n$ and thus a basis of $\mathbb{R}^n$ has n elements, no more no less.

We also note that if $\{a_1, ..., a_m\}$ is a linearly independent set in $\mathbb{R}^n$, then it can be extended to a basis $\{a_1, ..., a_m, a_{m+1}, .., a_n\}$ by adding suitable elements $a_{m+1}, .., a_n$. The extension starts by adding a_{m+1} as any vector which cannot be expressed as a linear combination of the set $\{a_1, ..., a_m\}$. Then $\{a_1, ..., a_m, a_{m+1}\}$ is linearly independent, and if $m + 1 < n$, the process may be continued.

We summarize as follows:

Theorem 42.1 *Any basis of $\mathbb{R}^n$ has n elements. Further, a set of n vectors in $\mathbb{R}^n$ span $\mathbb{R}^n$ if and only if it is linearly independent, that is a set of n vectors in $\mathbb{R}^n$ that spans $\mathbb{R}^n$ or is independent, must be a basis. Also, a set of fewer than n vectors in $\mathbb{R}^n$ cannot span $\mathbb{R}^n$, and a set of more than n vectors in $\mathbb{R}^n$ must be linearly dependent.*

The argument used to prove this result can also be used to prove that the dimension of a vector space S is well defined in the sense that any two bases have the same number of elements.

42.15 Coordinates in different bases

There are many different bases in $\mathbb{R}^n$ if $n > 1$ and the coordinates of a vector with respect to one basis are not equal to the coordinates with respect to another basis.

Suppose $\{a_1, ..., a_n\}$ is a basis for $\mathbb{R}^n$ and let us seek the connection between the coordinates of one and the same vector in the standard basis $\{e_1, ..., e_n\}$ and the basis $\{a_1, ..., a_n\}$. Assume then that the coordinates of the basis vectors a_j in the standard basis $\{e_1, ..., e_n\}$ are given by $a_j = (a_{1j}, ...a_{nj})$ for $j = 1, ...n$, that is

$$a_j = \sum_{i=1}^{n} a_{ij} e_i.$$

Denoting the coordinates of a vector x with respect to $\{e_1, ..., e_n\}$ by x_j and the coordinates with respect to $\{a_1, ..., a_n\}$ by $\hat{x}_j$, we have

$$x = \sum_{j=1}^{n} \hat{x}_j a_j = \sum_{j=1}^{n} \hat{x}_j \sum_{i=1}^{n} a_{ij} e_i = \sum_{i=1}^{n} (\sum_{j=1}^{n} a_{ij} \hat{x}_j) e_i, \tag{42.23}$$

that is since also $x = \sum_{i=1}^{n} x_i e_i$ and the coefficients x_i of x are unique,

$$x_i = \sum_{j=1}^{n} a_{ij} \hat{x}_j \quad \text{for } i = 1, ...n. \tag{42.24}$$

This relation expresses the connection between the coordinates $\hat{x}_j$ with respect to the basis $\{a_1, ..., a_n\}$, and the coordinates x_i with respect to the standard basis $\{e_1, ..., e_n\}$, in terms of the coordinates a_{ij} of the basis vectors a_j with respect to $\{e_1, ..., e_n\}$. This is a basic connection, which will play a central role in the development to come.

Using the scalar product we can express the coordinates a_{ij} of the basis vector a_j as $a_{ij} = (e_i, a_j)$. To find the connection (42.24) between the coordinates $\hat{x}_j$ with respect to the basis $\{a_1, ..., a_n\}$, and the coordinates x_i with respect to the standard basis $\{e_1, ..., e_n\}$, we may start from the equality $\sum_{j=1}^{n} x_j e_j = x = \sum_{j=1}^{n} \hat{x}_j a_j$ and take the scalar product of both sides with e_i, to get

$$x_i = \sum_{j=1}^{n} \hat{x}_j (e_i, a_j) = \sum_{j=1}^{n} a_{ij} \hat{x}_j, \tag{42.25}$$

where $a_{ij} = (e_i, a_j)$.

EXAMPLE 42.6. The set $\{a_1, a_2, a_3\}$ with $a_1 = (1, 0, 0)$, $a_2 = (1, 1, 0)$, $a_3 = (1, 1, 1)$ in the standard basis, forms a basis for $\mathbb{R}^3$ since the set $\{a_1, a_2, a_3\}$ is linearly independent. This is because if $\lambda_1 a_1 + \lambda_2 a_2 +$

$\lambda_3 a_3 = 0$, then $\lambda_3 = 0$ and thus also $\lambda_2 = 0$ and thus also $\lambda_1 = 0$. If (x_1, x_2, x_3) are the coordinates with respect to the standard basis and $(\hat{x}_1, \hat{x}_2, \hat{x}_3)$ are the coordinates with respect to $\{a_1, a_2, a_3\}$ of a certain vector, then the connection between the coordinates is given by $(x_1, x_2, x_3) = \hat{x}_1 a_1 + \hat{x}_2 a_2 + \hat{x}_3 a_3 = (\hat{x}_1 + \hat{x}_2 + \hat{x}_3, \hat{x}_2 + \hat{x}_3, \hat{x}_3)$. Solving for the $\hat{x}_j$ in terms of the x_i, we get $(\hat{x}_1, \hat{x}_2, \hat{x}_3) = (x_1 - x_2, x_2 - x_3, x_3)$.

42.16 Linear functions $f : \mathbb{R}^n \to \mathbb{R}$

A *linear function* $f : \mathbb{R}^n \to \mathbb{R}$ satisfies

$$f(x+y) = f(x) + f(y),\ f(\alpha x) = \alpha f(x) \quad \text{for all } x, y \in \mathbb{R}^n,\ \alpha \in \mathbb{R}. \tag{42.26}$$

We say that $f(x)$ is a *scalar linear function* since $f(x) \in \mathbb{R}$. Expressing $x = x_1 e_1 + ... + x_n e_n$ in the standard basis $\{e_1, ..., e_n\}$, and using the linearity of $f(x)$, we find that

$$f(x) = x_1 f(e_1) + ... + x_n f(e_n), \tag{42.27}$$

and thus $f(x)$ has the form

$$f(x) = f(x_1, ..., x_n) \quad = \quad a_1 x_1 + a_2 x_2 + ... + a_n x_n, \tag{42.28}$$

where the $a_j = f(e_j)$ are real numbers. We can write $f(x)$ as

$$f(x) = (a, x) = a \cdot x, \tag{42.29}$$

where $a = (a_1, ..., a_n) \in \mathbb{R}^n$, that is $f(x)$ can be expressed as the scalar product of x with the vector $a \in \mathbb{R}^n$ with components a_j given by $a_j = f(e_j)$.

The set of scalar linear functions is the mother of all other functions. We now generalize to systems of scalar linear functions. Linear algebra is the study of systems of linear functions.

EXAMPLE 42.7. $f(x) = 2x_1 + 3x_2 - 7x_3$ defines a linear function $f : \mathbb{R}^3 \to \mathbb{R}$ with coefficints $f(e_1) = a_1 = 2$, $f(e_2) = a_2 = 3$ and $f(e_3) = a_3 = -7$.

42.17 Linear transformations $f : \mathbb{R}^n \to \mathbb{R}^m$

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *linear* if

$$f(x+y) = f(x) + f(y),\ f(\alpha x) = \alpha f(x) \quad \text{for all } x, y \in \mathbb{R}^n,\ \alpha \in \mathbb{R}. \tag{42.30}$$

We also refer to a linear function $f : \mathbb{R}^n \to \mathbb{R}^m$ as a *linear transformation* of $\mathbb{R}^n$ into $\mathbb{R}^m$.

The image $f(x)$ of $x \in \mathbb{R}^n$ is a vector in $\mathbb{R}^m$ with components which we denote by $f_i(x)$, $i = 1, 2, ..., m$, so that $f(x) = (f_1(x), ..., f_m(x))$. Each *coordinate function* $f_i(x)$ is a linear scalar function $f_i : \mathbb{R}^n \to \mathbb{R}$ if $f : \mathbb{R}^n \to \mathbb{R}^m$ is linear. We can thus represent a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ as

$$\begin{array}{rcl} f_1(x) & = & a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n \\ f_2(x) & = & a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n \\ ...... & & \\ f_m(x) & = & a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n \end{array} \tag{42.31}$$

with the coefficients $a_{ij} = f_i(e_j) = (e_i, f(e_j)) \in \mathbb{R}$.

We can write (42.31) in condensed form as

$$f_i(x) = \sum_{j=1}^{n} a_{ij}x_j \quad \text{for } i = 1, ..., m. \tag{42.32}$$

EXAMPLE 42.8. $f(x) = (2x_1 + 3x_2 - 7x_3, x_1 + x_3)$ defines a linear function $f : \mathbb{R}^3 \to \mathbb{R}^2$ with coefficients $f_1(e_1) = a_{11} = 2$, $f_1(e_2) = a_{12} = 3$ and $f_1(e_3) = a_{13} = -7$, $f_2(e_1)a_{21} = 1$, $f_2(e_2)a_{22} = 0$ and $f_2(e_3) = a_{23} = 1$.

42.18 Matrices

We now return to the notion of a *matrix* and develop a matric calculus. The connection to linear transformations is very important. We define the $m \times n$ *matrix* $A = (a_{ij})$ as the rectangular array

$$\begin{pmatrix} a_{11} & a_{12} & .. & a_{1n} \\ . & . & .. & . \\ . & . & .. & . \\ a_{m1} & a_{m2} & .. & a_{mn} \end{pmatrix} \tag{42.33}$$

with rows $(a_{i1}, ..., a_{in})$, $i = 1, ..., m$, and columns $(a_{1j}, ..., a_{mj})$, $j = 1, ..., n$, where $a_{ij} \in \mathbb{R}$.

We may view each row $(a_{i1}, ..., a_{in})$ as a n-row vector or as a $1 \times n$ matrix, and each column $(a_{1j}, ..., a_{mj})$ as an m-column vector or a $m \times 1$ matrix. We can thus view the $m \times n$ matrix $A = (a_{ij})$ with elements a_{ij}, as consisting of m row vectors $(a_{i1}, ..., a_{in})$, $i = 1, ..., m$ or n column vectors $(a_{1j}, ..., a_{mj})$, $j = 1, ..., n$.

42.19 Matrix calculus

Let $A = (a_{ij})$ and $B = (b_{ij})$ be two $m \times n$ matrices. We define $C = A + B$ as the $m \times n$ matrix $C = (c_{ij})$ with elements

$$c_{ij} = a_{ij} + b_{ij}, \quad i = 1, ..., n, \, j = 1, ..., m. \tag{42.34}$$

We may thus add two $m \times n$ matrices by adding the corresponding elements.

Similarly, we define for λ a real number the matrix λA with elements (λa_{ij}), corresponding to multiplying all elements of A by the real number λ.

We shall now define matrix multiplication and we start by defining the product Ax of an $m \times n$ matrix $A = (a_{ij})$ with a $n \times 1$ column vector $x = (x_j)$ as the $m \times 1$ column vector $y = Ax$ with elements $y_i = (Ax)_i$ given by

$$(Ax)_i = \sum_{j=1}^{n} a_{ij} x_j, \tag{42.35}$$

or with matrix notation

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & & & \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}.$$

The element $y_i = (Ax)_i$ of the matrix-vector product Ax is thus obtained by taking the scalar product of row i of A with the vector x, as expressed by (42.35).

We can now express a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ as a matrix-vector product

$$f(x) = Ax,$$

where $A = (a_{ij})$ is an $m \times n$ matrix with elements $a_{ij} = f_i(e_j) = (e_i, f(e_j))$, where $f(x) = (f_1(x), ..., f_m(x))$. This is a restatement of (42.32).

We now proceed to define the product of an $m \times p$ matrix $A = (a_{ij})$ with a $p \times n$ matrix $B = (b_{ij})$. We do this by connecting the matrix product to the composition $f \circ g : \mathbb{R}^n \to \mathbb{R}^m$ given by

$$f \circ g(x) = f(g(x)) = f(Bx) = A(Bx), \tag{42.36}$$

where $f : \mathbb{R}^p \to \mathbb{R}^m$ is the linear transformation given by $f(y) = Ay$, where $A = (a_{ij})$ and $a_{ik} = f_i(e_k)$, and $g : \mathbb{R}^n \to \mathbb{R}^p$ is the linear transformation given by $g(x) = Bx$, where $B = (b_{kj})$ and $b_{kj} = g_k(e_j)$. Here e_k denote the standard basis vectors in $\mathbb{R}^p$, and e_j the corresponding basis vectors in $\mathbb{R}^n$. Clearly, $f \circ g : \mathbb{R}^n \to \mathbb{R}^m$ is linear and thus can be represented by an $m \times n$ matrix. Letting $(f \circ g)_i(x))$ denote the components of $(f \circ g)(x)$, we have

$$(f \circ g)_i(e_j) = f_i(g(e_j)) = f_i(\sum_{k=1}^{p} b_{kj} e_k) = \sum_{k=1}^{p} b_{kj} f_i(e_k) = \sum_{k=1}^{p} a_{ik} b_{kj},$$

which shows that $f \circ g(x) = Cx$, where $C = (c_{ij})$ is the $m \times n$ matrix with elements c_{ij} given by the formula

$$c_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}, \quad i = 1, ..., m,\ j = 1, ..., n. \tag{42.37}$$

We conclude that $A(Bx) = Cx$, and we are thus led to define the matrix product $AB = C$ by (42.37), where thus A is an $m \times p$ matrix and B is a $p \times n$ matrix, and the product AB is a $m \times n$ matrix. We can then write

$$A(Bx) = ABx$$

as a reflection of $f(g(x)) = f \circ g(x)$.

Formally, to get the $m \times n$ format of the product AB, we cancel the p in the $m \times p$ format of A and the $p \times n$ format of B. We see that the formula (42.37) may be expressed as follows: the element c_{ij} in row i and column j of AB is obtained by taking the scalar product of row i of A with column j of B.

We may write the formula for matrix multiplication as follows:

$$(AB)_{ij} = \sum_{k=1}^{p} a_{ik} b_{kj}, \quad \text{for } i = 1, .., n, j = 1, .., m, \tag{42.38}$$

or with matrix notation

$$AB = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \\ \dots & & & \\ a_{m1} & a_{m2} & \dots & a_{mp} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \dots & & & \\ b_{p1} & b_{p2} & \dots & b_{pn} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{k=1}^p a_{1k} b_{k1} & \sum_{k=1}^p a_{1k} b_{k2} & \dots & \sum_{k=1}^p a_{1k} b_{kn} \\ \sum_{k=1}^p a_{2k} b_{k1} & \sum_{k=1}^p a_{2k} b_{k2} & \dots & \sum_{k=1}^p a_{2k} b_{kn} \\ \dots & & & \\ \sum_{k=1}^p a_{mk} b_{k1} & \sum_{k=1}^p a_{mk} b_{k2} & \dots & \sum_{k=1}^p a_{mk} b_{kn} \end{pmatrix}.$$

Matrix multiplication is not commutative, that is $AB \neq BA$ in general. In particular, BA is defined only if $n = m$.

As a special case, we have that the product Ax of an $m \times n$ matrix A with an $n \times 1$ matrix x is given by (42.35). We may thus view the matrix-vector product Ax defined by (42.35) as a special case of the matrix product (42.37) with the $n \times 1$ matrix x being a column vector. The vector Ax is obtained taking the scalar product of the rows of A with the column vector x.

We sum up in the following theorem.

Theorem 42.2 *A linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ can be expressed as*

$$f(x) = Ax, \tag{42.39}$$

where $A = (a_{ij})$ is an $m \times n$ matrix with elements $a_{ij} = f_i(e_j) = (e_i, f(e_j))$, where $f(x) = (f_1(x), ..., f_m(x))$. If $g : \mathbb{R}^n \to \mathbb{R}^p$ and $f : \mathbb{R}^p \to \mathbb{R}^m$ are two linear transformations with corresponding matrices A and B, then the matrix of $f \circ g : \mathbb{R}^n \to \mathbb{R}^m$ is given by AB.

42.20 The transpose of a linear transformation

Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation defined by $f(x) = Ax$, where $A = (a_{ij})$ is an $m \times n$-matrix. We now define another linear transformation $f^\top : \mathbb{R}^m \to \mathbb{R}^n$, which we refer to as the *transpose* of f, by the relation

$$(x, f^\top(y)) = (f(x), y) \quad \text{for all } x \in \mathbb{R}^n,\, y \in \mathbb{R}^m. \tag{42.40}$$

Using that $f(x) = Ax$, we have

$$(f(x), y) = (Ax, y) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}x_j y_i, \tag{42.41}$$

and thus setting $x = e_j$, we see that

$$(f^\top(y))_j = \sum_{i=1}^{m} a_{ij}y_i. \tag{42.42}$$

This shows that $f^\top(y) = A^\top y$, where $A^\top$ is the $n \times m$ matrix with elements $(a^\top_{ji})$ given by $a^\top_{ji} = a_{ij}$. In other words, the columns of $A^\top$ are the rows of A and vice versa. For example, if

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}, \quad \text{then} \quad A^\top = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix}. \tag{42.43}$$

Summing up we have:

Theorem 42.3 *If $A = (a_{ij})$ is a $m \times n$ matrix, then the transpose $A^\top$ is an $n \times m$ matrix with elements $a^\top_{ji} = a_{ij}$, and*

$$(Ax, y) = (x, A^\top y) \quad \textit{for all } x \in \mathbb{R}^n,\, y \in \mathbb{R}^m. \tag{42.44}$$

An $n \times n$ matrix such that $A^\top = A$, that is $a_{ij} = a_{ji}$ for $i, j = 1, ...n$, is said to be a *symmetric* matrix.

42.21 Matrix norms

In many situations we need to estimate the "size" of a $m \times n$ matrix $A = (a_{ij})$. We may use this information to estimate the "length" of $y = Ax$ in terms of the "length" of x. We observe that

$$\sum_{i=1}^{m} |y_i| \leq \sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}||x_j| = \sum_{j=1}^{n}\sum_{i=1}^{m} |a_{ij}||x_j| \leq \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}| \sum_{j=1}^{n} |x_j|,$$

which shows that if we define $\|x\|_1 = \sum |x_j|$ and $\|y\|_1 = \sum |y_i|$, then

$$\|y\|_1 \leq \|A\|_1 \|x\|_1$$

if we define

$$\|A\|_1 = \max_{j=1,\dots,n} \sum_{i=1}^{m} |a_{ij}|$$

Similarly, we have

$$\max_i |y_i| \leq \max_i \sum_{j=1}^{n} |a_{ij}||x_j| \leq \max_i^n \sum_{j=1}^{n} |a_{ij}| \max_j |x_j|$$

which shows that if we define $\|x\|_\infty = \max_j |x_j|$ and $\|y\|_\infty = \max_i |y_i|$, then

$$\|y\|_\infty \leq \|A\|_\infty \|x\|_\infty$$

if we define

$$\|A\|_\infty = \max_{i+1,\dots,m} \sum_{j=1}^{n} |a_{ij}|.$$

We may also define the *Euclidean norm* $\|A\|$ by

$$\|A\| = \max_{x\in\mathbb{R}^n} \frac{\|Ax\|}{\|x\|}, \tag{42.45}$$

where we maximize over $x \neq 0$, and $\|\cdot\|$ denotes the Euclidean norm. We thus define $\|A\|$ to be the smallest constant C such that $\|Ax\| \leq C\|x\|$ for all $x \in \mathbb{R}^n$. We shall return in Chapter *The Spectral theorem* below to the problem of giving a formula for $\|A\|$ in terms of the coefficients of A in the case A is symmetric (with in particular $m = n$). By definition, we clearly have

$$\|Ax\| \leq \|A\|\,\|x\|. \tag{42.46}$$

If $A = (\lambda_i)$ is a diagonal $n \times n$ matrix with diagonal elements $a_{ii} = \lambda_i$, then

$$\|A\| = \max_{i=1,\dots,n} |\lambda_i|. \tag{42.47}$$

42.22 Lipschitz constant of a linear transformation

Consider a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ given by a $m \times n$ matrix $A = (a_{ij})$, that is

$$f(x) = Ax, \qquad \text{for } x \in \mathbb{R}^n. \tag{42.48}$$

By linearity we have

$$\|f(x) - f(y)\| = \|Ax - Ay\| = \|A(x - y)\| \leq \|A\|\|x - y\|.$$

We may thus say that the Lipschitz constant of $f : \mathbb{R}^n \to \mathbb{R}^m$ is equal to $\|A\|$. Alternatively, working with the norms $\|\cdot\|_1$ or $\|\cdot\|_\infty$, we may view the Lipschitz constant to be equal to $\|A\|_1$ or $\|A\|_\infty$.

42.23 Volume in $\mathbb{R}^n$. Determinant. Permutations

Let $\{a_1, a_2, ..., a_n\}$ be a set of n vectors in $\mathbb{R}^n$. We shall now generalize the concept of *volume* $V(a_1, ..., a_n)$ spanned by $\{a_1, a_2, ..., a_n\}$, which we have met above in the case $n = 2$ and $n = 3$. In particular, the volume will give us a tool to determine whether the set of vectors $\{a_1, .., a_n\}$ is linearly independent or not. Using the determinant we shall also develop Cramer's solution formula for an $n \times n$ system $Ax = b$, which generalizes the solution formulas for 2×2 and 3×3 which we have already met. The determinant is a quite complicated object, and we try to make the presentation as accessible as possible. When it comes to computing determinants. we shall return to the column echelon form.

Before actually giving a formula for the volume $V(a_1, ..., a_n)$ in terms of the coordinates $(a_{1j}, ..., a_{nj})$ of the vectors a_j, $j = 1, 2, ..., n$, we note that from our experience in $\mathbb{R}^2$ and $\mathbb{R}^3$, we expect $V(a_1, ..., a_n)$ to be a *multilinear alternating form*, that is

$$\begin{aligned} V(a_1, ..., a_n) &\in \mathbb{R}, \\ V(a_1, ..., a_n) &\quad \text{is linear in each argument } a_j, \\ V(a_1, ..., a_n) &= -V(\hat{a}_1, ..., \hat{a}_n), \end{aligned}$$

where $\hat{a}_1, ..., \hat{a}_n$ is a listing of $a_1, ..., a_n$ with two of the a_j interchanged. For example $\hat{a}_1 = a_2$, $\hat{a}_2 = a_1$ and $\hat{a}_j = a_j$ for $j = 3, ..., n$. We note that if two of the arguments in an alternating form is the same, for example $a_1 = a_2$, then $V(a_1, a_2, a_3, .., a_n) = 0$. This follows at once from the fact that $V(a_1, a_2, a_3, .., a_n) = -V(a_2, a_1, a_3, .., a_n)$. We are familiar with these properties in the case $n = 2, 3$.

We also need a little preliminary work on permutations. A *permutation* of the ordered list $\{1, 2, 3, 4, ..., n\}$ is a reordering of the list. For example $\{2, 1, 3, 4, ..., n\}$ is a permutation corresponding to interchanging the elements 1 and 2. Another permutation is $\{n, n-1, ..., 2, 1\}$ corresponding to reversing the order.

We can also describe a permutation as a one-to-one mapping of the set $\{1, 2, ..., n\}$ onto itself. We may denote the mapping by $\pi : \{1, 2, ..., n\} \to \{1, 2, ..., n\}$, that is $\pi(j)$ is one of the numbers $1, 2, ..., n$ for each $j = 1, 2, ..., n$ and $\pi(i) \neq \pi(j)$ if $i \neq j$. We can then talk about the *product* $\sigma\tau$ of two permutations σ and τ defined as the composition of τ and σ:

$$\sigma\tau(j) = \sigma(\tau(j)), \quad \text{for } j = 1, ..., n, \tag{42.49}$$

which is readily seen to be a permutation. Note that the order may be important: in general the permutation $\sigma\tau$ is different from the permutation $\tau\sigma$. In other words, multiplication of permutations is not commutative. However, multiplication is associative:

$$(\pi\sigma)\tau = \pi(\sigma\tau), \tag{42.50}$$

which directly follows from the definition by composition of functions.

A permutation corresponding to interchanging two elements, is called a *transposition*. More precisely, if π is a transposition then there are two elements p and q of the elements $\{1, 2, ..., n\}$, such that

$$\begin{aligned} \pi(p) &= q \\ \pi(q) &= p \\ \pi(j) &= j \quad \text{for } j \neq p,\, j \neq q. \end{aligned}$$

The permutation π with $\pi(j) = j$ for $j = 1, ..., n$ is called the identity permutation.

We shall use the following basic fact concerning permutations (a proof will be given in the Appendix).

Theorem 42.4 *Every permutation can be written as a product of transpositions. The representation is not unique, but for a given permutation the number of transpositions in such a representation cannot be odd in one case and even in another case; it is odd for all representations or even for all representations.*

We call a permutation *even* if it contains an even number of transposition factors, and *odd* if it contains an odd number of transpositions. The number of even perturbations is equal to the number of odd perturbations, and thus the total number of perturbations, including the identity, is even.

42.24 Definition of the volume $V(a_1, ...a_n)$

Assuming that $V(a_1, ..., a_n)$ is multilinear and alternating and that $V(e_1, e_2, ..., e_n) = 1$, we get the following relation

$$\begin{aligned} V(a_1, ..., a_n) &= V(\sum_j a_{j1}e_j, \sum_j a_{j2}e_j, ..., \sum_j a_{jn}e_j) \\ &= \sum_\pi \pm a_{\pi(1)\,1} a_{\pi(2)\,2} \cdots a_{\pi(n)\,n}, \end{aligned} \tag{42.51}$$

where we sum over all permutations π of the set $\{1, ..., n\}$, and the sign indicates if the permutation is even (+) or odd (-). Note that we give the identity permutation, which is included among the permutations, the sign +. We recall that $a_j = (a_{1j}, ..., a_{nj})$ for $j = 1, ..., n$.

We now turn around in this game, and simply take (42.51) as a definition of the volume $V(a_1, ..., a_n)$ spanned by the set of vectors $\{a_1, ..., a_n\}$. By this definition it follows that $V(a_1, ..., a_n)$ is indeed a multilinear alternating form on $\mathbb{R}^n$. Further, $V(e_1, ..., e_n) = 1$, since the only non-zero term in the sum (42.51) in this case corresponds to the identity perturbation.

We can transform the definition $V(a_1, ..., a_n)$ to matrix language as follows. Let $A = (a_{ij})$ be the $n \times n$ matrix

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ . & . & \dots & . \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \tag{42.52}$$

formed by the column vectors $a_1, ..., a_n$ with coefficients $a_j = (a_{1j}, ...a_{nj})$. We define the *determinant* det A of A, by

$$\det A = V(a_1, ..., a_n) = \sum_\pi \pm a_{\pi(1)\,1} a_{\pi(2)\,2} \cdots a_{\pi(n)\,n},$$

where we sum over all permutations π of the set $\{1, ..., n\}$, and the sign indicates if the permutation is even (+) or odd (-).

We note that since the unit vectors e_j in $\mathbb{R}^n$ are mapped by A into the column vectors a_j, that is since $Ae_j = a_j$, we have that A maps the unit n-cube in $\mathbb{R}^n$ onto the parallelepiped in $\mathbb{R}^n$ spanned by a_1,..., a_n. Since the volume of the n-cube is one and the volume of the the parallelepiped spanned by a_1,..., a_n is $V(a_1, ...a_n)$, the *volume scale* of the mapping $x \to Ax$ is equal to $V(a_1, ..., a_n)$.

42.25 The volume $V(a_1, a_2)$ in $\mathbb{R}^2$

If A is the 2×2-matrix

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$

then det $A = V(a_1, a_2)$ is given

$$\det A = V(a_1, a_2) = a_{11}a_{22} - a_{21}a_{12}. \tag{42.53}$$

of course, $a_1 = (a_{11}, a_{21})$ and $a_2 = (a_{12}, a_{22})$ are the column vectors of A.

42.26 The volume $V(a_1, a_2, a_3)$ in $\mathbb{R}^3$

If A is the 3×3-matrix

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

then det $A = V(a_1, a_2, a_3)$ is given by

$$\begin{aligned} &\det A = V(a_1, a_2, a_3) = a_1 \cdot a_2 \times a_3 \\ &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}). \end{aligned} \tag{42.54}$$

We see that we can express det A as

$$\begin{aligned}\det A &= a_{11}\det A_{11} - a_{12}\det A_{12} + a_{13}A_{13}\\ &= a_{11}V(\hat{a}_2,\hat{a}_3) - a_{12}V(\hat{a}_1,\hat{a}_3) + a_{13}V(\hat{a}_1,\hat{a}_2) \quad (42.55)\end{aligned}$$

where the A_{1j} are 2×2 matrices formed by cutting out the first row and j:th column of A, explicitely given by

$$A_{11} = \begin{pmatrix} a_{22} & a_{23}\\ a_{32} & a_{33}\end{pmatrix} \quad A_{12} = \begin{pmatrix} a_{21} & a_{23}\\ a_{31} & a_{33}\end{pmatrix} \quad A_{13} = \begin{pmatrix} a_{21} & a_{22}\\ a_{31} & a_{32}\end{pmatrix}$$

and $\hat{a}_1 = (a_{21}, a_{31})$, $\hat{a}_2 = (a_{22}, a_{32})$, $\hat{a}_3 = (a_{23}, a_{33})$ are the 2-column vectors formed by cutting out the first element of the 3-columns a_j. We say that (42.55) is an *expansion* of the 3×3 matrix A in terms of the elements of the first row of A and the corresponding 2×2 matrices. The expansion formula follows by collecting all the terms with a_{11} as a factor, and all the terms with a_{12} as a factor and all the terms with a_{13} as a factor.

42.27 The volume $V(a_1, a_2, a_3, a_4)$ in $\mathbb{R}^4$

Using the expansion formula we can compute the determinant det $A = V(a_1, ..., a_4)$ of a 4×4 matrix $A = (a_{ij})$ with column vectors $a_j = (a_{1j}, ..., a_{4j})$ for $j = 1, 2, 3, 4$. We have

$$\begin{aligned}\det A = V(a_1, a_2, a_3, a_4) &= a_{11}V(\hat{a}_2,\hat{a}_3,\hat{a}_4) - a_{12}V(\hat{a}_1,\hat{a}_3,\hat{a}_4)\\ &+ a_{13}V(\hat{a}_1,\hat{a}_2,\hat{a}_4) - a_{14}V(\hat{a}_1,\hat{a}_2,\hat{a}_3),\end{aligned}$$

where the $\hat{a}_j$, $j = 1, 2, 3, 4$ are the 3-column vectors corresponding to cutting out the first coefficient of the a_j. We have now expressed the determinant of the 4×4 matrix A as a sum of determinants of 3×3 matrices with the first row of A as coefficients.

42.28 The volume $V(a_1, ..., a_n)$ in $\mathbb{R}^n$

Iterating the row-expansion formula indicated above, we can compute the determinant of an arbitrary $n\times n$ matrix A. As an example we give the expansion formula for a 5×5 matrix $A = (a_{ij})$:

$$\begin{aligned}\det A = V(a_1, a_2, a_3, a_4, a_5) &= a_{11}V(\hat{a}_2,\hat{a}_3,\hat{a}_4,\hat{a}_5) - a_{12}V(\hat{a}_1,\hat{a}_3,\hat{a}_4,\hat{a}_5)\\ +a_{13}V(\hat{a}_1,\hat{a}_2,\hat{a}_4,\hat{a}_5) &- a_{14}V(\hat{a}_1,\hat{a}_2,\hat{a}_3,\hat{a}_5) + a_{15}V(\hat{a}_1,\hat{a}_2,\hat{a}_3,\hat{a}_4).\end{aligned}$$

Evidently, we can formulate the following a rule of sign for the term with the factor a_{ij}: choose the $+$ if $i + j$ is even and the $-$ if $i + j$ is odd. This rule generalizes to expansions with respect to any row of A.

42.29 The determinant of a triangular matrix

Let $A = (a_{ij})$ be a *upper triangular* $n \times n$ matrix, that is $a_{ij} = 0$ for $i > j$. All elements a_{ij} of A below the diagonal are zero. In this case the only non-zero term in the expression for $\det A$, is the product of the diagonal elements of A corresponding to the identity perturbation, so that

$$\det A = a_{11}a_{22}\cdots a_{nn}. \tag{42.56}$$

This formula also applies to a *lower triangular* $n \times n$ matrix $A = (a_{ij})$ with $a_{ij} = 0$ for $i < j$.

42.30 Computing $\det A = V(a_1, ..., a_n)$ by reduction to column echelon form

We now present a way to compute $\det A = V(a_1, ..., a_n)$, where the a_j are the columns of a $n \times n$ matrix $A = (a_{ij})$, based on reduction to column echelon form. We then use that the volume does not change if we subtract one column multiplied by a real number from another column, to obtain

$$\det A = V(a_1, a_2, ..., a_n) = V(\hat{a}_1, \hat{a}_2, \hat{a}_3, ..., \hat{a}_n)$$

where $\hat{a}_{ij} = 0$ if $j > i$, that is the corresponding matrix $\hat{A}$ is a lower triangular matrix. We then compute $V(\hat{a}_1, \hat{a}_2, \hat{a}_3, ..., \hat{a}_n)$ by multiplying the diagonal elements. As usual, if we meet a zero diagonal term we interchange columns until we meet a nonzero diagonal term, or if all diagonal terms appearing this way are zero, we proceed to modify the next row. At least one of the diagonal terms in the final trianguar matrix will then be zero, and thus the determinant will be zero.

EXAMPLE 42.9. We show the sequence of matrices in a concrete case:

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 4 & 6 \\ 3 & 4 & 6 \end{pmatrix} \quad \to \begin{pmatrix} 1 & 0 & 0 \\ 2 & 2 & 4 \\ 3 & 1 & 3 \end{pmatrix} \quad \to \begin{pmatrix} 1 & 0 & 0 \\ 2 & 2 & 0 \\ 3 & 1 & 1 \end{pmatrix}$$

and conclude that *det* $A = 2$.

42.31 The magic formula $\det AB = \det A \det B$

Let A and B be two $n \times n$ matrices. We know that AB is the matrix of the composite transformation $f(g(x))$, where $f(y) = Ay$ and $g(x) = Bx$. The volume scale of the mapping $x \to Bx$ is equal to $\det B$ and the volume scale of the mapping $y \to Ay$ is $\det A$, and hence the volume scale of the mapping $x \to ABx$ is equal to $\det A \det B$. This proves that

$$\det AB = \det A \det B,$$

which is one of the corner stones of the calculus of determinants. The proof suggested is a "short proof" avoiding algebraic computations. One can also give a direct algebraic proof using suitable expansion formulas for the determinant.

42.32 Test of linear independence

To test the linear independence of a given set of n vectors $\{a_1, a_2, ..., a_n\}$ in $\mathbb{R}^n$, we can use the volume $V(a_1, a_2, ..., a_n)$. More precisely, we shall prove that $\{a_1, a_2, ..., a_n\}$ is linearly independent if and only if $V(a_1, a_2, ..., a_n) \neq 0$. First, we note that if $\{a_1, a_2, ..., a_n\}$ is linearly dependent, for example if $a_1 = \sum_{j=2}^n \lambda_j a_j$ is a linear combination of $\{a_2, ..., a_n\}$, then $V(a_1, a_2, ..., a_n) = \sum_{j=2}^n \lambda_j V(a_j, a_2, ..., a_n) = 0$, since each factor $V(a_j, a_2, ..., a_n)$ has two equal vectors.

Secondly, if $\{a_1, a_2, ..., a_n\}$ is linearly independent, that is if $\{a_1, a_2, ..., a_n\}$ is a basis for $\mathbb{R}^n$, then we must have $V(a_1, ..., a_n) \neq 0$. We see this as follows. We express each e_j as a linear combination of the set $\{a_1, a_2, ..., a_n\}$, for example $e_1 = \sum \lambda_{1j} a_j$. We have, since V is multilinear and vanishes if two arguments are the same, and $V(a_{\pi(1)}, .., a_{\pi(n)}) = \pm V(a_1, .., a_n)$ for any permutation π, that

$$\begin{aligned} 1 = V(e_1, .., e_n) &= V(\sum_j \lambda_{1j} a_j, e_2, .., e_n) = \sum_j \lambda_{1j} V(a_j, e_2, .., e_n) \\ &= \sum_j \lambda_{1j} V(a_j, \sum_k \lambda_{2k} a_k, e_3, .., e_n) = ... = cV(a_1, .., a_n), \end{aligned} \tag{42.57}$$

for some constant c. We conclude that $V(a_1, ..., a_n) \neq 0$. We summarize as follows:

Theorem 42.5 *A set $\{a_1, a_2, ..., a_n\}$ of n vectors in $\mathbb{R}^n$ is linearly independent if and only if $V(a_1, ..., a_n) \neq 0$.*

We may restate this result in matrix language as follows: The columns of an $n \times n$-matrix A are linearly independent if and only if $\det A \neq 0$. We may thus sum up as follows:

Theorem 42.6 *Let A be a $n \times n$ matrix. Then the following statements are equivalent:*

- *The columns of A are linearly independent.*

- *If* $Ax = 0$ *then* $x = 0$.

- $\det A \neq 0$.

To test linear independence of the columns of a given matrix A we may thus compute $\det A$ and check if $\det A = 0$. We can also use this test in more quantitative form as follows: If $\det A$ is small then the columns are close to being linearly dependent and uniqueness is at risk!

A matrix A with $\det A = 0$ is called *singular*, while matrices with $\det A \neq 0$ are refered to as *non-singular*. Thus an $n \times n$-matrix is non-singular if and only if its columns are linearly independent. Again we can go to quantitative forms and say that a matrix A is close to singular if its determinant is close to zero. The dependence of the solution on the size of the determinant is clearly expressed in the next chapter.

42.33 Cramer's solution formula for non-singular $n \times n$ systems.

Consider again the $n \times n$ linear system of equations

$$Ax = b \tag{42.58}$$

or

$$\sum_{j=1}^{n} a_j x_j = b \tag{42.59}$$

where $A = (a_{ij})$ is an $n \times n$ matrix with columns $a_j = (a_{1j}, ..., a_{nj})$, $j = 1, ..., n$. Suppose that the columns a_j of A are linearly independent, or equivalently, that $\det A = V(a_1, ..., a_n) \neq 0$. We then know that (42.58) has a unique solution $x \in \mathbb{R}^n$ for any given $b \in \mathbb{R}^n$, we shall now seek a formula for the solution x in terms of b and the columns a_j of A.

Using the basic property of the volume function $V(g_1, ..., g_n)$ of a set $\{g_1, ..., g_n\}$ of n vectors g_i, in particular the property that $V(g_1, ..., g_n) = 0$ if any two of the g_i are equal, we obtain the following solution formula (Cramer's formula):

$$\begin{aligned} x_1 &= \frac{V(b, a_2, ..., a_n)}{V(a_1, a_2, ..., a_n)}, \\ &.... \\ x_n &= \frac{V(a_1, ..., a_{n-1}, b)}{V(a_1, a_2, ..., a_n)}. \end{aligned} \tag{42.60}$$

For example, to obtain the formula for x_1, use that

$$V(b, a_2, ..., a_n) = V(\sum_j a_j x_j, a_2, ..., a_n)$$

$$= \sum_{j=1}^{n} x_j V(a_j, a_2, ..., a_n) = x_1 V(a_1, a_2, .., a_n).$$

We summarize:

Theorem 42.7 *If A is a $n \times n$ non-singular matrix with* $\det A \neq 0$*, then the system of equations $Ax = b$ has a unique solution x for any $b \in \mathbb{R}^n$. The solution is given by Cramer's formula* (42.60).

A result like this was first derived by Leibniz and then by Gabriel Cramer (1704-1752) (who got a Ph D at the age of 18 with a thesis on the theory of sound) in *Introduction l'analyse des lignes courbes algbraique.* Throughout the book Cramer makes essentially no use of the Calculus in either Leibniz' or Newton's form, although he deals with such topics as tangents, maxima and minima, and curvature, and cites Maclaurin and Taylor in footnotes. One conjectures that he never accepted or mastered Calculus.

FIGURE 42.2. Gabriel Cramer: "I am friendly, good-humoured, pleasant in voice and appearance, and possess good memory, judgement and health".

Note that Cramer's solution formula for $Ax = b$ is very computationally demanding, and thus cannot be used for actually computing the solution unless n is small. To solve linear systems of equations other methods are used, like Gaussian elimination and iterative methods, see Chapter *Solving systems of linaer equations.*

42.34 Inverse matrix

Let A be a non-singular $n \times n$ matrix with $V(a_1, ..., a_n) \neq 0$. Then $Ax = b$ can be solved uniquely for all $b \in \mathbb{R}^n$ according to Cramer's solution formula (42.60). Clearly, x depends linearly on b, and the solution x may be expressed as $A^{-1}b$, where A^{-1} is an $n \times n$ matrix which we refer to as the *inverse* of A. The j:th column of A^{-1} is the solution vector corrsponding to choosing $b = e_j$. Cramer's formula thus gives the following formula for the inverse A^{-1} of A:

$$A^{-1} = V(a_1, .., a_n)^{-1} \begin{pmatrix} V(e_1, a_2, .., a_n) & .. & V(a_1, .., a_{n-1}, e_1) \\ . & .. & . \\ . & .. & . \\ V(e_n, a_2, .., a_n) & .. & V(a_1, .., a_{n-1}, e_n) \end{pmatrix}. \tag{42.61}$$

The inverse matrix A^{-1} of A satisfies

$$A^{-1}A = AA^{-1} = I,$$

where I is the $n \times n$ identity matrix, with ones on the diagonal and zeros elsewhere.

Evidently, we can express the solution to $Ax = b$ in the form $x = A^{-1}b$ if A is a non-singular $n \times n$ matrix (by multiplying $Ax = b$ by A^{-1} from the left)..

42.35 Projection onto a subspace

Let V be a subspace of $\mathbb{R}^n$ spanned by the linearly independent set of vectors $\{a_1, ..., a_m\}$. In other words, $\{a_1, ..., a_m\}$ is a basis for V. The projection Pv of a vector $v \in \mathbb{R}^n$ onto V is defined as the vector $Pv \in V$ satisfying the orthogonality relation

$$(v - Pv, w) = 0 \quad \text{for all vectors } w \in V, \tag{42.62}$$

or equivalently

$$(Pv, a_j) = (v, a_j) \quad \text{for } j = 1, .., m. \tag{42.63}$$

To see the equivalence, we note that (42.62) clearly implies (42.63). Conversely, any $w \in V$ is a linear combination of the form $\sum \mu_j a_j$, and multiplying (42.63) by μ_j and summing over j, we obtain $(Pv, \sum_j \mu_j a_j) = (v, \sum_j \mu_j a_j)$, which is (42.62) with $w = \sum_j \mu_j a_j$ as desired.

Expressing $Pv = \sum_{i=1}^m \lambda_i a_i$ in the basis $\{a_1, ..., a_m\}$ for V, the orthogonality relation (42.63) corresponds to the $m \times m$ linear system of equations

$$\sum_{i=1}^m \lambda_i (a_i, a_j) = (v, a_j) \quad \text{for } j = 1, 2, ...m. \tag{42.64}$$

We shall now prove that this system admits a unique solution, which proves that the projection Pv of v onto V exists and is unique. By Theorem **??** it is enough to prove uniqueness. We thus assume that

$$\sum_{i=1}^{m} \lambda_i (a_i, a_j) = 0 \quad \text{for } j = 1, 2, ...m.$$

Multiplying by λ_j and summing we get

$$0 = (\sum_{i=1}^{m} \lambda_i a_i, \sum_{j=1}^{m} \lambda_j a_j) = |\sum_{i=1}^{m} \lambda_i a_i|^2,$$

which proves that $\sum_i \lambda_i a_i = 0$ and thus $\lambda_i = 0$ for $i = 1, ..., m$, since the $\{a_1, ..., a_m\}$ is linearly independent.

We have now proved the following fundamental result:

Theorem 42.8 *Let V be a linear subspace of $\mathbb{R}^n$. Then for all $v \in \mathbb{R}^n$ the projection Pv of v onto V, defined by $Pv \in V$ and $(v - Pv, w) = 0$ for all $w \in V$, exists and is unique.*

We note that $P : \mathbb{R}^n \to V$ is a linear mapping. To see this, let v and $\hat{v}$ be two vectors in $\mathbb{R}^n$, and note that since $(v - Pv, w) = 0$ and $(\hat{v} - P\hat{v}, w) = 0$ for all $w \in V$, we have

$$(v + \hat{v} - (Pv + P\hat{v}), w) = (v - Pv, w) + (\hat{v} - P\hat{v}, w) = 0,$$

which shows that $Pv + P\hat{v} = P(v + \hat{v})$. Similarly, $Pw = \lambda Pv$ if $w = \lambda v$, for any $\lambda \in \mathbb{R}$ and $v \in \mathbb{R}^n$. This proves the linearity of $P : \mathbb{R}^n \to V$.

We further note that $PP = P$. We sum up as follows:

Theorem 42.9 *The projection $P : \mathbb{R}^n \to V$ onto a linear subspace V of $\mathbb{R}^n$ is a linear transformation defined by $(v - Pv, w) = 0$ for all $w \in V$, which satisfies $PP = P$.*

42.36 An equivalent characterization of the projection

We shall now prove that the *projection* Pv of a vector $v \in \mathbb{R}^n$ onto V is the vector $Pv \in V$ with minimum distance to v, that is $|v - Pv| \leq |v - w|$ for all $w \in V$.

We state the equivalence of the two definitions of the projection in the following fundamental theorem:

Theorem 42.10 . *Let $v \in \mathbb{R}^n$ be given. The vector $Pv \in V$ satisfies the orthogonality relation*

$$(v - Pv, w) = 0 \quad \textit{for all vectors } w \in V, \tag{42.65}$$

if and only if Pv minimizes the distance to v in the sense that

$$|v - Pv| \leq |v - w| \quad \textit{for all } w \in V. \tag{42.66}$$

Further, the element $Pv \in V$ satisfying (42.65) *and* (42.66) *is uniquely determined.*

To prove the theorem we note that by the orthogonality (42.62), we have for any $w \in V$,

$$\begin{aligned}|v - Pv|^2 &= (v - Pv, v - Pv)\\ &= (v - Pv, v - w) + (v - Pv, w - Pv) = (v - Pv, v - w),\end{aligned}$$

since $w - Pv \in V$. Using Cauchy-Schwarz inequality, we obtain

$$|v - Pv|^2 \leq |v - Pv|\,|v - w|,$$

which shows that $|v - Pv| \leq |v - w|$ for all $w \in V$.

Conversely, if $|v - Pv| \leq |v - w|$ for all $w \in V$, then for all $\epsilon \in \mathbb{R}$ and $w \in V$

$$\begin{aligned}|v - Pv|^2 &\leq |v - Pv + \epsilon w|^2\\ &= |v - Pv|^2 + \epsilon(v - Pv, w) + \epsilon^2|w|^2,\end{aligned}$$

that is for all $\epsilon > 0$

$$(v - Pv, w) + \epsilon|w|^2 \geq 0,$$

which proves that

$$(v - Pv, w) \geq 0 \quad \text{for all } w \in V.$$

Changing w to $-w$ proves the reverse inequality and we conclude that $(v - Pv, w) = 0$ for all $w \in V$.

Finally, to prove uniqueness, assume that $z \in V$ satisfies

$$(v - z, w) = 0 \quad \text{for all vectors } w \in V.$$

Then $(Pv - z, w) = (Pv - v, w) + (v - z, w) = 0 + 0 = 0$ for all $w \in V$, and $Pv - z$ is a vector in V. Choosing $w = Pv - z$ thus shows that $|Pv - z|^2 = 0$, that is $z = Pv$. The proof of the theorem is now complete.

The argument just given is very fundamental and will be used many times below in various forms, so it is worth taking the time to understand it now.

42.37 Orthogonal decomposition. Pythagoras theorem

Let V be a subspace of $\mathbb{R}^n$. Let P be the projection onto V. Any vector x can be written

$$x = Px + (x - Px) \tag{42.67}$$

where $Px \in V$, and further $(x - Px) \perp V$ since by the definition of P we have $(x - Px, w) = 0$ for all $w \in V$. We say that $x = Px + (x - Px)$ is an *orthogonal decomposition* of x since $(Px, x - Px) = 0$.

Define the *orthogonal complement* $V^\perp$ to V by $V^\perp = \{y \in \mathbb{R}^n : y \perp V\} = \{y \in \mathbb{R}^n : y \perp x \quad \text{for all } x \in V\}$. It is clear that $V^\perp$ is a linear subspace of $\mathbb{R}^n$. We have that if $x \in V$ and $y \in V^\perp$, then $(x, y) = 0$. Further, any vector $z \in \mathbb{R}^n$ can be written $z = x + y$, with $x = Pz \in V$ and $y = (x - Px) \in V^\perp$. We can summarize by saying that

$$V \oplus V^\perp = \mathbb{R}^n, \tag{42.68}$$

is an *orthogonal decomposition* of $\mathbb{R}^n$ into the two orthogonal subspaces V and $V^\perp$: $x \in V$ and $y \in V^\perp$ implies $(x, y) = 0$ and any $z \in \mathbb{R}^n$ can be written uniquely in the form $z = x+y$. The uniqueness of the decomposition $z = Pz + (z - Pz)$ follows from the uniqueness of Pz.

We note the following generalization of Pythagoras theorem: for any $x \in \mathbb{R}^n$, we have

$$|x|^2 = |Px|^2 + |x - Px|^2. \tag{42.69}$$

This follows by writing $x = Px + (x - Px)$ and using that $Px \perp (x - Px)$:

$$|x|^2 = |Px + (x - Px)|^2 = |Px|^2 + 2(Px, x - Px) + |x - Px|^2.$$

More generally, we have if $z = x + y$ with $x \perp y$ (that is $(x, y) = 0$), that

$$|z|^2 = |x|^2 + |y|^2.$$

42.38 Properties of projections

Let P be the orthogonal projection onto a linear subspace V in $\mathbb{R}^n$. Then $P : \mathbb{R}^n \to \mathbb{R}^n$ is a linear transformation that satisfies

$$P^\top = P \quad \text{and} \quad PP = P. \tag{42.70}$$

We have already seen that $PP = P$. To see that $P^\top = P$ we note that

$$(w, P^\top v) = (Pw, v) = (Pw, Pv) = (w, Pv) \quad \text{for all } v, w \in \mathbb{R}^n, \tag{42.71}$$

and thus $P^\top = P$. Conversely, let $P : \mathbb{R}^n \to \mathbb{R}^n$ be a linear transformation which satisfies (42.70). Then P is the orthogonal projection onto a subspace V of $\mathbb{R}^n$. To see this, set $V = R(P)$ and note that since $P^\top = P$ and $PP = P$, we have

$$\begin{aligned}(x - Px, Px) &= (x, Px) - (Px, Px) = (x, Px) - (x, P^\top Px)\\ &= (x, Px) - (x, Px) = 0.\end{aligned}$$

This shows that $x = Px + (x - Px)$ is an orthogonal decomposition, and thus P is the orthogonal projection onto $V = R(P)$.

42.39 Orthogonalization: the Gram-Schmidt process

Let $\{a_1, ..., a_m\}$ be a basis for a subspace V of $\mathbb{R}^n$. In other words, $\{a_1, ..., a_m\}$ is linearly independent and V is the set of linear combinations of $\{a_1, ..., a_m\}$. We try to construct another basis $\{\hat{e}_1, ..., \hat{e}_m\}$ for V that is *orthonormal*, i.e. such that the basis vectors $\hat{e}_i$ are mutually orthogonal and have length equal to one or

$$(\hat{e}_i, \hat{e}_j) = 0 \quad \text{for } i \neq j, \quad \text{and } |\hat{e}_i| = 1 \tag{42.72}$$

We choose $\hat{e}_1 = \frac{1}{|a_1|}a_1$ and let V_1 be the subspace spanned by $\hat{e}_1$, or equivalently by a_1. Let P_1 be the projection onto V_1. Define

$$\hat{e}_2 = \frac{1}{|a_2 - P_1 a_2|}(a_2 - P_1 a_2).$$

Then $(\hat{e}_1, \hat{e}_2) = 0$ and $|\hat{e}_2| = 1$. Further, the subspace V_2 spanned by $\{a_1, a_2\}$ is also spanned by $\{\hat{e}_1, \hat{e}_2\}$. We now continue in the same way: Let P_2 be the projection onto V_2 and define

$$\hat{e}_3 = \frac{1}{|a_3 - P_2 a_3|}(a_3 - P_2 a_3)$$

Then the subspace V_3 spanned by $\{a_1, a_2, a_3\}$ is also spanned by the orthonormal set $\{\hat{e}_1, \hat{e}_2, \hat{e}_3\}$.

Continuing, we obtain an orthonormal basis $\{\hat{e}_1, ..., \hat{e}_m\}$ for the subspace spanned by $\{a_1, ..., a_m\}$ with the property that for $i = 1, ..., m$, the subspace spanned by $\{a_1, ..., a_i\}$ is spanned by $\{\hat{e}_1, ..., \hat{e}_i\}$.

Note that since the basis $\{\hat{e}_1, ..., \hat{e}_m\}$ is orthogonal, the system of equations (42.64) corresponding to computing $P_{i-1}a_i$, is diagonal.

42.40 Orthogonal matrix

Consider the matrix Q with columns $\hat{e}_1, .., \hat{e}_n$, where $\{\hat{e}_1, .., \hat{e}_n\}$ is an orthonormal basis for $\mathbb{R}^n$. Since the vectors $\hat{e}_j$ are pairwise orthogonal and of length one, $Q^\top Q = I$, where I is the $n \times n$ identity matrix. Conversely, if Q is a matrix such that $Q^\top Q = I$, where I is an identity matrix, then the columns of Q must be orthonormal.

An $n \times n$-matrix Q such that $Q^\top Q = I$, is called an *orthogonal matrix*. An orthogonal $n \times n$-matrix can thus be characterized as follows: Its columns form an *orthonormal basis* for $\mathbb{R}^n$, that is a basis consisting of pairwise orthogonal vectors of length, or norm, one.

We summarize:

Theorem 42.11 *An orthogonal matrix Q satisfies $Q^\top Q = QQ^\top = I$, and $Q^{-1} = Q^\top$.*

42.41 Invariance of the scalar product under orthogonal transformations

Let Q by an $n \times n$ orthonormal matrix with the columns formed by the coefficients of basis vectors $\hat{e}_j$ of an orthonormal basis $\{\hat{e}_1, .., \hat{e}_n\}$. We then know that the coordinates x of a vector with respect to the standard basis, and the coordinates $\hat{x}$ with respect to the basis $\{\hat{e}_1, .., \hat{e}_n\}$, are connected by

$$x = Q\hat{x}.$$

We now prove that the scalar product is invariant under the orthonormal change of coordinates $x = Q\hat{x}$. We compute setting $y = Q\hat{y}$,

$$(x, y) = (Q\hat{x}, Q\hat{y}) = (Q^\top Q\hat{x}, \hat{y}) = (\hat{x}, \hat{y}),$$

that is the scalar product is the same in the $\{e_1, .., e_n\}$ coordinates as in the $\{\hat{e}_1, .., \hat{e}_n\}$ coordinates. We summarize:

Theorem 42.12 *If Q is an orthogonal $n \times n$ matrix, then $(x, y) = (Qx, Qy)$ for all $x, y \in \mathbb{R}^n$.*

42.42 The QR-decomposition

We can give the Gram-Schmidt orthogonalization process the following matrix interpretation: Let $\{a_1, ..., a_m\}$ be m linearly independent vectors in $\mathbb{R}^n$ and let A be the $n \times m$ matrix with the a_j occurring as columns. Let $\{\hat{e}_1, ..., \hat{e}_m\}$ be the corresponding orthonormal set generated by the Gram-Schmidt process, and let Q be the $n \times m$ matrix with the $\hat{e}_j$ as columns. Then

$$A = QR, \tag{42.73}$$

where R is a $m \times m$ upper triangular matrix, which expresses each a_j as a linear combination of $\{\hat{e}_1, ..., \hat{e}_j\}$.

The matrix Q satisfies $Q^\top Q = I$, where I is the $m \times m$ identity matrix, since the $\hat{e}_j$ are pairwise orthogonal and have length 1. We conclude that a $m \times n$ matrix A with linearly independent columns can be factored into $A = QR$, where Q satisfies $Q^\top Q = I$, and R is upper triangular. The columns of the matrix Q are orthonormal, as in the case of an orthonormal matrix, but if $m < n$, then they do not span all of $\mathbb{R}^n$.

42.43 The Fundamental Theorem of Linear Algebra

We return to our basic question of existence and uniqueness of solutions to the system $Ax = b$ with A a given $m \times n$ matrix and $b \in \mathbb{R}^m$ a given vector.

We now allow m to be different from n, remembering that we focussed on the case $m = n$ above. We shall now prove the Fundamental Theorem of Linear Algebra giving an answer of theoretical nature to our basic questions of existence and uniqueness.

We note the following chain of equivalent statements for a $m \times n$-matrix A, where "iff" is shorthand for "if and only if":

$$\begin{aligned} x \in N(A) \quad \text{iff} \quad Ax = 0 \quad \text{iff} \quad x \perp \text{ rows of } A \quad \text{iff} \\ x \perp \text{ columns of } A^\top \quad \text{iff} \\ x \perp R(A^\top) \quad \text{iff} \\ x \in (R(A^\top)^\perp. \end{aligned}$$

Thus $N(A) = (R(A^\top)^\perp$, and since $R(A^\top)^\perp \oplus R(A^\top) = \mathbb{R}^n$, we see that

$$N(A) \oplus R(A^\top) = \mathbb{R}^n. \tag{42.74}$$

As a consequence of this orthogonal splitting, we see that

$$\dim N(A) + \dim R(A^\top) = n, \tag{42.75}$$

where dim V is the dimension of the linear space V. We recall that the dimension dim V of a linear space V is the number of elements in a basis for V. Similarly, replacing A by $A^\top$ and using that $(A^\top)^\top = A$, we have

$$N(A^\top) \oplus R(A) = \mathbb{R}^m, \tag{42.76}$$

and thus in particular,

$$\dim N(A^\top) + \dim R(A) = m. \tag{42.77}$$

Next we note that, letting $g_1, ..., g_k$ be a basis in $N(A)^\perp$ so that $Ag_1, ..., Ag_k$ span $R(A)$ and thus dim $R(A) \leq k$, we have

$$\dim N(A) + \dim R(A) \leq n, \quad \text{and also} \quad \dim N(A^\top) + \dim R(A^\top) \leq m. \tag{42.78}$$

Adding (42.75) and (42.77), we conclude that equality holds in (42.78). We summarize in:

Theorem 42.13 (The Fundamental Theorem of Linear Algebra) *Let A be a $m \times n$ matrix. Then*

$$N(A) \oplus R(A^\top) = \mathbb{R}^n \quad N(A^\top) \oplus R(A) = \mathbb{R}^m,$$

$$dim\ N(A) + dim\ R(A^\top) = n,\ dim\ N(A^\top) + dim\ R(A) = m,$$

$$dim\ N(A) + dim\ R(A) = n,\ dim\ N(A^\top) + dim\ R(A^\top) = m,$$

$$dim\ R(A) = dim\ R(A^\top).$$

In the special case $m = n$, we have that $R(A) = \mathbb{R}^m$ if and only if $N(A) = 0$ (which we proved above using Cramer's rule), stating that uniqueness implies existence.

We call dim $R(A)$ the *column rank* of the matrix A. The column rank of A is equal to the dimension of the space spanned by the columns of A. Similarly the *row rank* of A is equal to the dimension of the space spanned by the rows of A. The equality dim $R(A) =$ dim $R(A^\top)$ in the Fundamental Theorem expresses that the the column ranks of A and $A^\top$ are equal, that is that the column rank of A is equal to the *row rank* of A. We state this result as:

Theorem 42.14 *The number of linearly independent columns of A is equal to the number of linearly independent rows of A.*

EXAMPLE 42.10. Returning to Example 41.5, we note that the column echelon form of $A^\top$ is the transpose of the row echelon form of A, that is

$$R(A^\top) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 1 & 3 & 2 & 0 \\ 1 & 6 & 5 & 0 \end{pmatrix}.$$

We check that the columns vectors $(0, 1, -2, 1, 0)$ and $(0, 4, -5, 0, 1)$ spanning $N(A)$ are orthogonal to $R(A^\top)$, that is orthogonal to the columns of the echelon form of $A^\top$. Of course, this is just a restatement of the fact that these vectors are orthogonal to the rows of the row echelon form $\hat{A}$ of A (as is evident from the proof of the Fundamental Theorem). We see that $N(A) \oplus R(A^\top) = \mathbb{R}^5$ as predicted by the Fundamental Theorem.

42.44 Change of basis: coordinates and matrix representations

Let $\{s_1, ..., s_n\}$ be a basis for $\mathbb{R}^n$ where the coordinates of the basis vectors in the standard basis $\{e_1, ..., e_n\}$ are given by $s_j = (s_{1j}, ..., s_{nj})$. Recalling (42.24), we have the following connection between the coordinates x_i of a vector x with respect to the standard basis and the coordinates $\hat{x}_j$ of x with respect to the basis $\{s_1, ..., s_n\}$:

$$x_i = \sum_{j=1}^{n} s_{ij}\hat{x}_j \quad \text{for } i = 1, ...n. \tag{42.79}$$

This follows from taking the scalar product of $\sum_{j=1}^{n} x_j e_j = \sum_{i=1}^{n} \hat{x}_j s_j$ with e_i and using that $s_{ij} = (e_i, s_j)$.

Introducing the matrix $S = (s_{ij})$, we thus have the following connection between the coordinates $x = (x_1, ..., x_n)$ with respect to $\{e_1, ..., e_n\}$, and the coordinates $\hat{x} = (\hat{x}_1, ...\hat{x}_n)$ with respect to $\{s_1, ..., s_n\}$:

$$x = S\hat{x}, \quad \text{that is } \hat{x} = S^{-1}x. \tag{42.80}$$

Consider now a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^n$ with matrix $A = (a_{ij})$ with respect to the standard basis $\{e_1, ...e_n\}$, that is with $a_{ij} = f_i(e_j) = (e_i, f(e_j))$, where $f(x) = (f_1(x), ..., f_n(x))$ in the standard basis $\{e_1, ...e_n\}$, that is

$$y = f(x) = \sum_i f_i(x)e_i = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_je_i = Ax.$$

Writing $y = S\hat{y}$ and $x = S\hat{x}$, we have

$$S\hat{y} = AS\hat{x} \quad \text{that is } \hat{y} = S^{-1}AS\hat{x}$$

This shows that the matrix of the linear transformation $f : \mathbb{R}^n \to \mathbb{R}^n$, with the matrix A with respect to the standard basis, takes the following form in the basis $\{s_1, ...s_n\}$:

$$S^{-1}AS, \tag{42.81}$$

where the the coefficients s_{ij} of the matrix $S = (s_{ij})$ are the coordinates of the basis vectors s_j with respect to the standard basis.

42.45 Least squares methods

Consider the $m \times n$ linear system of equations

$$Ax = b,$$

or

$$\sum_j^n a_jx_j = b,$$

where $A = (a_{ij})$ is an $m \times n$ matrix with columns $a_j = (a_{1j}, ..., a_{mj})$, $j = 1, ..., n$. We know that if $b \in R(A)$ then the system can be solved, and if $N(A) = 0$, then the solution is unique. Suppose now that $b \notin R(A)$. Then there is no $x \in \mathbb{R}^n$ such that $Ax = b$, and the system $Ax = b$ has no solution. We can replace the problem by the following *least squares problem*

$$\min_{x\in\mathbb{R}^n} |Ax - b|^2.$$

This problem amounts to seeking the projection Pb of b onto $R(A)$, that is the projection of b onto the space spanned by the columns a_j of A.

By the properties of projections given above, we know that $Pb \in R(A)$ exists and is uniquely determined by the relation

$$(Pb, y) = (b, y) \quad \text{for all } y \in R(A),$$

that is we seek $Pb = A\hat{x}$ for some $\hat{x} \in \mathbb{R}^n$ such that

$$(A\hat{x}, Ax) = (b, Ax) \quad \text{for all } x \in \mathbb{R}^n.$$

This relation can be written

$$(A^\top A\hat{x}, x) = (A^\top b, x) \quad \text{for all } x \in \mathbb{R}^n,$$

which is the same as the matrix equation

$$A^\top A\hat{x} = A^\top b,$$

which we refer to as the *normal equations.*

The matrix $A^\top A$ is an $n \times n$ symmetric matrix. Assume now that the columns a_j of A are linearly independent. Then $A^\top A$ is non-singular, because if $A^\top Ax = 0$, then

$$0 = (A^\top Ax, x) = (Ax, Ax) = |Ax|^2,$$

and thus $Ax = 0$ and therefore $x = 0$, since the columns of A are linearly independent. Thus the equation $A^\top A\hat{x} = A^\top b$ has a unique solution $\hat{x}$ for each right hand side $A^\top b$, given by the formula

$$\hat{x} = (A^\top A)^{-1}A^\top b.$$

In particular, we have the following formula for the projection Pb of b onto $R(A)$,

$$Pb = A(A^\top A)^{-1}A^\top b.$$

We can directly check that $P : \mathbb{R}^m \to \mathbb{R}^m$ defined this way is symmetric and satisfies $P^2 = P$.

If the columns of A are linearly dependent, then $\hat{x}$ is undetermined up to vectors $\hat{x}$ in $N(A)$. It is then natural to single out a unique $\hat{x}$ by requiring that $|\hat{x}|^2$ to be minimal. Using the orthogonal decomposition $\mathbb{R}^n = R(A^\top) \oplus N(A)$, this is equivalent to seeking $\hat{x}$ in $R(A^\top)$, since by Pythagoras theorem this minimizes $|\hat{x}|$. We thus seek $\hat{x}$ so that

- $A\hat{x}$ is equal to the projection Pb of b onto $R(A)$
- $\hat{x} \in R(A^\top)$.

This leads to the following equation for $\hat{x} = A^\top \hat{y}$:

$$(A\hat{x}, AA^\top y) = (b, AA^\top y) \quad \text{for all } y \in \mathbb{R}^m, \tag{42.82}$$

with $\hat{x}$ uniquely determined.

Chapter 42 Problems

42.1. Prove that a plane in $\mathbb{R}^3$ not passing through the origin is not a subspace of $\mathbb{R}^3$.

42.2. (a) What is a vector space? (b) What is a subspace of a vector space?

42.3. Verify (42.17) and (42.18).

42.4. Why must a set of more than n vectors in $\mathbb{R}^n$ be linearly dependent? Why must a set of n linearly independent vectors in $\mathbb{R}^n$ be a basis?

42.5. Verify that $R(A)$ and $N(A)$ are linear subspaces of $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively, and further that the orthogonal complement $V^\top$ of a subspace V of $\mathbb{R}^n$ is also a subspace of $\mathbb{R}^n$.

42.6. (a) Give an example showing that permutations need not commute. (b) Verify the associative law for permutations.

42.7. Compute the determinants of some $n \times n$ matrices with $n = 2, 3, 4, 5$.

42.8. Fill out the details in the proof of Cauchys inequality.

42.9. Write an algorithm for the Gram-Schmidt orthogonalization process, and implement it in Matlab, for example.

42.10. Fill in the details in (42.57)

42.11. Verify that for an orthogonal matrix $QQ^\top = I$. Hint: Multiply $Q^\top Q = I$ from the right with C and from the right with Q, where C is the matrix such that $QC = I$.

42.12. Prove for 2×2 matrices A and B that $\det AB = \det A \det B$.

42.13. How many operations are needed to solve an $n \times n$ system of linear equations using Cramer's formula?

42.14. Prove by reduction to column echelon form that a basis for $\mathbb{R}^n$ contains n elements.

42.15. Implement algorithms for reduction to column and row echelon forms.

42.16. Prove that the solution $\hat{x} \in R(A^\top)$ of (42.82) is uniquely determined.

42.17. Construct the row and column echelon forms of different (small) matrices, and check the validity of the Fundamental Theorem.

43
The Spectral Theorem

There seems to be three possibilities (of a Unified Theory of Physics):

1. There really is a complete unified theory, which we will someday discover if we are smart enough.
2. There is no ultimate theory of the Universe, just an infinite sequence of theories that describe the Universe more and more accurately.
3. There is no theory of the Universe; events cannot be predicted beyond a certain extent but occur in a random and arbitray manner. (Stephen Hawking, in *A Brief History of Time*)

43.1 Eigenvalues and eigenvectors

Let $A = (a_{ij})$ be a quadratic $n \times n$ matrix. We investigate the situation in which multiplication by A acts like scalar multiplication. To start with, we assume that the elements a_{ij} are real numbers. If $x = (x_1, ..., x_n) \in \mathbb{R}^n$ is a non-zero vector that satisfies

$$Ax = \lambda x, \tag{43.1}$$

where λ is a real number, then we say that $x \in \mathbb{R}^n$ is an *eigenvector* of A and that λ is a corresponding *eigenvalue* of A. An eigenvector x has the property that Ax is parallel to x (if $\lambda \neq 0$), or $Ax = 0$ (if $\lambda = 0$). This is a special property, as easy to verify with almost any example we might make up.

If x is an eigenvector with corresponding eigenvalue λ then $\bar{x} = \mu x$ for any non-zero real number μ is also an eigenvector corresponding to the eigenvalue λ because

$$\text{if } Ax = \lambda x, \text{ then } A\bar{x} = \mu Ax = \mu\lambda x = \lambda\mu x = \lambda\bar{x}.$$

Thus, we may change the length of an eigenvector without changing the corresponding eigenvalue. For example, we may normalize an eigenvector to have length equal to 1. In essence, the direction of an eigenvector is determined, but not its length.

We shall now study the problem of finding eigenvalues and corresponding eigenvectors of a given a quadratic matrix. We shall see that this is a basic problem of linear algebra arising in many different situations. We shall prove the *Spectral Theorem* stating that if A is a symmetric real $n \times n$ matrix, then there is an orthogonal basis for $\mathbb{R}^n$ consisting of eigenvectors. We shall also briefly discuss the case of non-symmetric matrices.

Rewriting (43.1) as $(A - \lambda I)x = 0$ with $x \in \mathbb{R}^n$ a non-zero eigenvector and I the identity matrix, we see that the matrix $A - \lambda I$ must be singular if λ is an eigenvalue, that is $\det(A - \lambda I) = 0$. Conversely, if $\det(A - \lambda I) = 0$ then $A - \lambda I$ is singular and thus the null-space $N(A - \lambda I)$ is different from the zero vector and thus there is a non-zero vector x such that $(A - \lambda I)x = 0$, that is there is an eigenvector x with corresponding eigenvalue λ. Using the expansion formula for the determinant, we see that $\det(A - \lambda I)$ is a polynomial in λ of degree n with coefficients depending on the coefficients a_{ij} of A. The polynomial equation

$$\det(A - \lambda I) = 0$$

is called the *characteristic equation*. We summarize:

Theorem 43.1 *The number λ is an eigenvalue of the $n \times n$ matrix A if and only if λ is a root of the characteristic equation* $\det(A - \lambda I) = 0$.

EXAMPLE 43.1.

If $A = (a_{ij})$ is a 2×2 matrix, then the characteristic equation is

$$\det(A - \lambda I) = (a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} = 0,$$

which is a second order polynomial equation in λ. For example, if

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

then the characteristic equation is $\det(A - \lambda I) = \lambda^2 - 1 = 0$ with roots $\lambda_1 = 1$ and $\lambda_2 = -1$. The corresponding normalized eigenvectors are $s_1 = \frac{1}{\sqrt{2}}(1, 1)$ and $s_2 = \frac{1}{\sqrt{2}}(1, -1)$ since

$$(A - \lambda_1 I)\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

and similarly $(A-\lambda_2)s_2 = 0$. We observe that $(s_1, s_2) = s_1 \cdot s_2 = 0$, that is eigenvectors corresponding to different eigenvalues are orthogonal.

43.2 Basis of eigenvectors

Suppose $\{s_1, ..., s_n\}$ is a basis for $\mathbb{R}^n$ consisting of eigenvectors of the $n \times n$ matrix $A = (a_{ij})$ with corresponding eigenvalues λ_1,...,λ_n so

$$As_i = \lambda_i s_i \quad \text{for } i = 1, ..., n. \tag{43.2}$$

Let S be the matrix with the columns equal to the eigenvectors s_j expressed in the standard basis. We can then write (43.2) in matrix form as follows,

$$AS = SD, \tag{43.3}$$

where D is the diagonal matrix with the eigenvalues λ_j on the diagonal. We thus have

$$A = SDS^{-1} \qquad \text{or} \qquad D = S^{-1}AS, \tag{43.4}$$

where D is a diagonal matrix. We say that S transforms A into a diagonal matrix D with the eigenvalues on the diagonal.

Conversely, if we can express a matrix A in the form $A = SDS^{-1}$ with S non-singular and D diagonal then $AS = SD$, which says that the columns of S are eigenvectors with corresponding eigenvalues on the diagonal of D.

Viewing the $n \times n$ matrix A as defining a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^n$ by $f(x) = Ax$, we can express the action of $f(x)$ in a basis of eigenvectors $\{s_1, ..., s_n\}$ by the diagonal matrix D since $f(s_i) = \lambda_i s_i$. Thus, the linear transformation $f : \mathbb{R}^n \to \mathbb{R}^n$ is expressed by the matrix A in the standard basis and by the diagonal matrix D matrix in a basis of eigenvectors. The coupling is given by

$$D = S^{-1}AS.$$

Of course, the action of a diagonal matrix is very easy to describe and to understand and this is the motivation for considering eigenvalues and eigenvectors.

We now formulate the following basic question in two equivalent forms:

- Given a $n \times n$ matrix A, is there a basis of eigenvectors of A?
- Given a $n \times n$ matrix A, is there a non-singular matrix S such that $S^{-1}AS$ is diagonal?

As we have seen, the columns of the matrix S are the eigenvectors of A and the diagonal elements are the eigenvalues.

We shall now give the following partial answer: if A is an $n \times n$ symmetric matrix, then there is an orthogonal basis for $\mathbb{R}^n$ consisting of eigenvectors.

This is the celebrated *Spectral Theorem for symmetric matrices.* Notice the assumption that A is symmetric and that in this case the basis of eigenvectors may be chosen to be orthogonal.

EXAMPLE 43.2. Recalling Example 43.1, we see that $s_1 = \frac{1}{\sqrt{2}}(1,1)$ and $s_2 = \frac{1}{\sqrt{2}}(1,-1)$ form an orthogonal basis. By the orthogonality of S, $S^{-1} = S^\top$, and

$$S^{-1}AS = \frac{1}{2}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}$$
$$= \frac{1}{2}\begin{pmatrix}1 & 1\\ -1 & 1\end{pmatrix}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix} = \begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}.$$

43.3 The Spectral Theorem for symmetric matrices: easy version

The following version of the Spectral Theorem for symmetric matrices is easy to prove:

Theorem 43.2 *Let A be a symmetric $n \times n$ matrix. Suppose A has n distinct eigenvalues $\lambda_1,\dots, \lambda_n$ and corresponding normalized eigenvectors $s_1,\dots,s_n$ with $\|s_j\| = 1$, $j = 1,\dots,n$. Then, $\{s_1,\dots,s_n\}$ is an orthonormal basis of eigenvectors. Letting $Q = (q_{ij})$ be the orthogonal matrix with the columns $(q_{1j},\dots,q_{nj})$ being the coordinates of the eigenvectors g_j with respect to the standard basis, then $D = Q^{-1}AQ$ is a diagonal matrix with the eigenvalues λ_j on the diagonal and $A = QDQ^{-1}$, where $Q^{-1} = Q^\top$.*

To prove this result, it suffices to prove that eigenvectors corresponding to different eigenvalues are orthogonal. This follows from the assumption that there are n distinct eigenvalues $\lambda_1,\dots, \lambda_n$ with corresponding normalized eigenvectors $s_1,\dots,s_n$. If we prove that these eigenvectors are pairwise orthogonal, then they form a basis for $\mathbb{R}^n$ and the proof is complete. Thus, assume that s_i and s_j are eigenvectors corresponding to different eigenvalues λ_i and λ_j. Since A is symmetric and $(Ax,y) = (x,Ay)$ for all $x,y \in \mathbb{R}^n$, we have

$$\lambda_i(s_i,s_j) = (\lambda_i s_i,s_j) = (As_i,s_j) = (s_i,As_j)$$
$$= (s_i,\lambda_j s_j) = \lambda_j(s_i,s_j),$$

which implies that $(s_i,s_j) = 0$ since $\lambda_i \neq \lambda_j$. We state this observation as a theorem because of its basic importance.

Theorem 43.3 *If A is a symmetric $n\times n$ matrix, and s_i and s_j are eigenvectors of A corresponding to the eigenvalues λ_i and λ_j with $\lambda_i \neq \lambda_j$, then*

$(s_i, s_j) = 0$. *In other words, eigenvectors corresponding to different eigenvalues are orthogonal.*

Note that above we prove the Spectral Theorem for a symmetric $n \times n$ matrix A in the case the characteristic equation $\det(A - \lambda I) = 0$ has n different roots. It thus remains to consider the case of multiple roots where there are less than n different roots. We will consider this below. The reader in hurry may skip that proof.

43.4 Applying the Spectral Theorem to an IVP

We show a typical application of the Spectral Theorem. Consider the initial value problem: find $u : [0,1] \to \mathbb{R}^n$ such that

$$\dot{u} = Au, \quad \text{for } 0 < t \leq 1, \quad u(0) = u_0,$$

where $A = (a_{ij})$ is a symmetric $n \times n$ matrix with real coefficients a_{ij} independent of t. Systems of this form arise in many applications and the behavior of such a system may be very complicated.

Suppose now that $\{g_1, ..., g_n\}$ is an orthonormal basis of eigenvectors of A and let Q be the matrix with columns comprised of the coordinates of the eigenvectors g_j with respect to the standard basis. Then $A = QDQ^{-1}$, where D is the diagonal matrix with the eigenvalues λ_j on the diagonal. We introduce the new variable $v = Q^{-1}u$, that is we set $u = Qv$, where $v : [0,1] \to \mathbb{R}^n$. Then, the equation $\dot{u} = Au$ takes the form $Q\dot{v} = AQv$, that is $\dot{v} = Q^{-1}AQv = Dv$, where we use the fact that Q is independent of time. Summing up, we get the following diagonal system in the new variable v,

$$\dot{v} = Dv \quad \text{for } 0 < t \leq 1, \quad v(0) = v_0 = Q^{-1}u_0.$$

The solution of this decoupled system is given by

$$v(t) = \begin{pmatrix} \exp(\lambda_1 t) & 0 & 0 & .... & 0 \\ 0 & \exp(\lambda_2 t) & 0 & .... & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & .... & \exp(\lambda_n) \end{pmatrix} v_0 = \exp(Dt)v_0,$$

where $\exp(Dt)$ is a diagonal matrix with the elements $\exp(\lambda_j t)$ on the diagonal. The dynamics of this system is easy to grasp: each component $v_j(t)$ of $v(t)$ evolves according to $v_j(t) = \exp(\lambda_j t)v_{0j}$.

Transforming back, we get the following solution formula in the original variable $u(t)$,

$$u(t) = Q\exp(Dt)Q^{-1}u_0. \tag{43.5}$$

With A as in (43.1), we get the solution formula

$$u(t) = \frac{1}{2}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\begin{pmatrix}e^t & 0\\ 0 & e^{-t}\end{pmatrix}\begin{pmatrix}1 & 1\\ 1 & -1\end{pmatrix}\begin{pmatrix}u_{01}\\ u_{02}\end{pmatrix}$$
$$= \frac{1}{2}\begin{pmatrix}(e^t + e^{-t})u_{01} + (e^t - e^{-t})u_{02}\\ (e^t - e^{-t})u_{01} + (e^t + e^{-t})u_{02}\end{pmatrix}.$$

43.5 The Spectral Theorem for symmetric matrices: general case

Above, we saw that eigenvalues of a matrix A are roots of the characteristic equation $\det(A - \lambda I) = 0$. In principle, we can find the eigenvalues and eigenvectors of given matrix by first solving the characteristic equation to find all the eigenvalues, and then for each eigenvalue λ find corresponding eigenvectors by solving the linear system of equations $(A - \lambda I)x = 0$.

We shall now present an alternative way of constructing/finding the eigenvectors and eigenvalues of a symmetric matrix A that also proves the Spectral Theorem for a symmetric $n \times n$ matrix A in the general case with possibly multiple roots. In the proof, we construct an orthonormal basis of eigenvectors $\{s_1, ..., s_n\}$ of A by constructing the eigenvectors one by one starting with s_1.

Constructing the first eigenvector s_1

To construct the first eigenvector s_1, we consider the minimization problem: find $\bar{x} \in \mathbb{R}^n$ such that

$$F(\bar{x}) = \min_{x \in \mathbb{R}^n} F(x), \tag{43.6}$$

where

$$F(x) = \frac{(Ax, x)}{(x, x)} = \frac{(f(x), x)}{(x, x)} \tag{43.7}$$

is the so-called *Rayleigh quotient*. We note that the function $F(x)$ is *homogenous of degree zero*, that is for any $\lambda \in \mathbb{R}$, $\lambda \neq 0$, we have

$$F(x) = F(\lambda x),$$

because we can simply divide out the factor λ. In particular, for any $x \neq 0$,

$$F(x) = F(\frac{x}{\|x\|}), \tag{43.8}$$

and thus we may restrict the x in (54.13) to have length one, that is we may consider the equivalent minimization problem: find $\bar{x}$ with $\|\bar{x}\| = 1$ such that

$$F(\bar{x}) = \min_{x \in \mathbb{R}^n \, \|x\|=1} F(x) \tag{43.9}$$

Since $F(x)$ is Lipschitz continuous on the closed and bounded subset $\{x \in \mathbb{R}^n : \|x\| = 1\}$ of $\mathbb{R}^n$, we know by the Chapter Minimization, that the problem (43.9) has a solution $\bar{x}$, and thus also the problem (54.13) has a solution $\bar{x}$. We set $s_1 = \bar{x}$, and check that g_1 is indeed an eigenvector of A, that is an eigenvector of $f : \mathbb{R}^n \to \mathbb{R}^n$.

Since $\bar{x}$ solves the minimization problem (54.13), we have $\nabla F(\bar{x}) = 0$, where ∇F is the gradient of F. Computing $\nabla F(x)$ using the symmetry of $F(x)$ or the matrix A, we find that

$$\nabla F(x) = \frac{(x,x)2Ax - (Ax,x)2x}{(x,x)^2}, \tag{43.10}$$

so that with $x = \bar{x}$ satisfying $(\bar{x}, \bar{x}) = 1$,

$$\nabla F(\bar{x}) = 2(A\bar{x} - (A\bar{x}, \bar{x})\bar{x}) = 0,$$

that is

$$A\bar{x} = \lambda_1 \bar{x}, \tag{43.11}$$

where

$$\lambda_1 = (A\bar{x}, \bar{x}) = \frac{(A\bar{x}, \bar{x})}{(\bar{x}, \bar{x})} = \min_{x \in \mathbb{R}^n} F(x). \tag{43.12}$$

Setting $s_1 = \bar{x}$, we thus have

$$As_1 = \lambda_1 s_1, \quad \lambda_1 = (As_1, s_1), \quad \|s_1\| = 1.$$

We have now constructed the first normalized eigenvector s_1 with corresponding eigenvalue λ_1. We now let V_1 be the orthogonal complement of the space spanned by s_1, consisting of all the vectors $x \in \mathbb{R}^n$ such that $(x, s_1) = 0$. The dimension of V_1 is $n - 1$.

Invariance of A

Note that V_1 is *invariant* with respect to A in the sense that if $x \in V_1$ then $Ax \in V_1$. This follows because if $(x, g_1) = 0$ then $(Ax, s_1) = (x, As_1) = (x, \lambda_1 s_1) = \lambda_1 (x, s_1) = 0$. This means that we can now restrict attention to the action of A on V_1, having handled the action of A on the space spanned by the first eigenvector s_1.

Constructing the second eigenvector s_2

Consider the minimization problem to find $\bar{x} \in V_1$ such that

$$F(\bar{x}) = \min_{x \in V_1} F(x). \tag{43.13}$$

By the same argument, this problem has a solution which we denote s_2 and which satisfies $As_2 = \lambda_2 s_2$, where $\lambda_2 = \frac{(As_2, s_2)}{(s_2, s_2)}$, and $\|s_2\| = 1$. Because in

(43.13) we minimize over a smaller set than in (54.13), $\lambda_2 \geq \lambda_1$. Note that it may happen that $\lambda_2 = \lambda_1$, although V_1 is a subset of $\mathbb{R}^n$. In that case, we say that $\lambda_1 = \lambda_2$ is a *multiple eigenvalue.*

Continuing the process

Let V_2 be the orthogonal subspace to the space spanned by s_1 and s_2 Again A is invariant on V_2 and the space spanned by $\{s_1, s_2\}$. Continuing this way, we obtain a orthonormal basis $\{s_1, ..., s_n\}$ of eigenvectors of A with corresponding real eigenvalues λ_i.

We have now proved the famous

Theorem 43.4 (Spectral Theorem): *If $f : \mathbb{R} \to \mathbb{R}$ is a linear symmetric transformation with corresponding symmetric $n \times n$ matrix A in the standard basis, then there is an orthogonal basis $(g_1, ..., g_n)$ of $\mathbb{R}^n$ consisting of eigenvectors g_i of f with corresponding real eigenvalues λ_j satisfying $f(g_j) = Ag_j = \lambda_j g_j$, for $j = 1, ..., n$. We have $D = Q^{-1}AQ$ and $A = QDQ^{-1}$, where Q is the orthogonal matrix with the coefficients of the eigenvectors g_j in the standard basis forming the columns, and D is the diagonal matrix with the eigenvalues λ_j on the diagonal.*

43.6 The norm of a symmetric matrix

We recall that we have defined the *Euclidean norm* $\|A\|$ of a $n \times n$ matrix A by

$$\|A\| = \max_{x \in \mathbb{R}^n} \frac{\|Ax\|}{\|x\|}, \tag{43.14}$$

where we maximize over $x \neq 0$. By the definition, we have

$$\|Ax\| \leq \|A\| \, \|x\|, \tag{43.15}$$

and we may thus view $\|A\|$ to be the smallest constant C such that $\|Ax\| \leq C\|x\|$ for all $x \in \mathbb{R}^n$.

We shall now prove that if A is symmetric, then we can directly relate $\|A\|$ to the eigenvalues $\lambda_1, ..., \lambda_n$ of A:

$$\|A\| = \max_{i=1,\ldots,n} |\lambda_i|. \tag{43.16}$$

We do this as follows. Using the Spectral theorem, we can write A as $A = Q^\top \Lambda Q$ with Q orthogonal and Λ a diagonal matrix with the eigenvalues λ_i on the diagonal. We recall that (cf. (42.47))

$$\|\Lambda\| = \max_{i=1,\ldots,n} |\lambda_i| = |\lambda_j| \tag{43.17}$$

and thus for all $x \in \mathbb{R}^n$,

$$\|Ax\| = \|Q^\top \Lambda Qx\| = \|\Lambda Qx\| \leq \|\Lambda\|\|Qx\| = \|\Lambda\|\|x\| = \max_{i=1,\ldots,n} |\lambda_i|\|x\|,$$

which proves that $\|A\| \leq \max_{i=1,\ldots,n} |\lambda_i|$. Choosing x to be equal to the eigenvector corresponding to the eigenvalue λ_j of maximal modulus proves that indeed $\|A\| = \max_{i=1,\ldots,n} |\lambda_i| = |\lambda_j|$. We have proved the following result, which is a corner stone of numerical linear algebra.

Theorem 43.5 *If A is a symmetric $n \times n$ matrix, then $\|A\| = \max_{i=1,\ldots,n} |\lambda_i|$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A.*

43.7 Extension to non-symmetric real matrices

Up until now, we have mainly focussed on the case of *real scalars*, that is we assume that the components of vectors are real numbers. We know that we can also let the components of vectors be *complex numbers*, and we may then allow eigenvalues to be complex numbers. The fundamental theorem of algebra states that a polynomial equation of degree n with complex coefficients, has n complex roots, and thus the characteristic equation $\det(A - \lambda I) = 0$ has n complex roots $\lambda_1, \ldots, \lambda_n$, and thus a $n \times n$ matrix A has n complex eigenvalues $\lambda_1, \ldots, \lambda_n$, if roots are counted with multiplicity. We have in this chapter focussed on *symmetric* matrices A with real coefficients and we have proved that a symmetric matrix with real coefficients has n real eigenvalues, counted with multiplicity. For symmetric matrices we can thus limit ourselves to real roots of the characteristic equation.

Chapter 43 Problems

43.1. Verify (43.10).

43.2. Compute the eigenvalues and eigenvetors of the following matrices: (a)... (b)...

This is page 657
Printer: Opaque this

44
Solving linear algebraic systems

All thought is a kind of computation. (Hobbes)

44.1 Introduction

We are interested in solving a system of linear equations

$$Ax = b,$$

where A is a given $n \times n$ matrix and $b \in \mathbb{R}^n$ is a given n-vector and we seek the solution vector $x \in \mathbb{R}^n$. We recall that if A is non-singular with non-zero determinant, then the solution $x \in \mathbb{R}^n$ is theoretically given by Cramer's formula. However if n is large, the computational work in using Cramer's formula is prohibitively large, so we need to find a more efficient means of computing the solution.

We shall consider two types of methods for solving the system $Ax = b$: (i) *direct methods* based on *Gaussian elimination* and theoretically produce a solution after a finite number of arithmetic operations, and (ii) *iterative methods* that produce a generally infinite sequence of increasingly accurate approximations. We consider in particular iterative methods for A symmetric and positive-definite.

44.2 Direct methods

We begin by noting that some linear systems are easier to solve than others. For example if $A = (a_{ij})$ is *diagonal*, which means that $a_{ij} = 0$ if $i \neq j$, then the system is solved in n operations: $x_i = b_i/a_{ii}$, $i = 1, ..., n$. Further, if the matrix is *upper triangular*, which means that $a_{ij} = 0$ if $i > j$, or *lower triangular*, which means that $a_{ij} = 0$ if $i < j$, then the system can be solved by *backward substitution* or *forward substitution* respectively; see Fig. 44.1 for an illustration of these different types. For example if A is

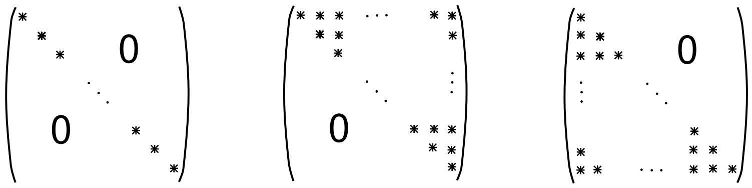


FIGURE 44.1. The pattern of entries in diagonal, upper, and lower triangular matrices. A "$*$" denotes a possibly nonzero entry.

upper triangular, the "pseudo-code" shown in Fig. 44.2 solves the system $Ax = b$ for the vector $x = (x_i)$ given the vector $b = (b_i)$ (assuming that $a_{kk} \neq 0$): In all three cases, the systems have a unique solution as long as

```
x_n = b_n/a_nn
for k = n-1, n-2, ..., 1, do
    sum = 0
    for j = k+1, ..., n, do
        sum = sum + a_kj · x_j
    x_k = (b_k - sum)/a_kk
```

FIGURE 44.2. An algorithm for solving an upper triangular system by back substitution.

the diagonal entries of A are nonzero.

Direct methods are based on Gaussian elimination, which in turn is based on the observation that the solution of a linear system is not changed under the following *elementary row operations*:

- interchanging two equations

- adding a multiple of one equation to another
- multiplying an equation by a nonzero constant.

The idea behind Gaussian elimination is to transform using these operations a given system into an upper triangular system, which is solved by back substitution. For example, to solve the system

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_2 + 2x_3 &= 1 \\ 2x_1 + x_2 + 3x_3 &= 1, \end{aligned}$$

we first subtract 2 times the first equation from the third to get the equivalent system,

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_2 + 2x_3 &= 1 \\ -x_2 + x_3 &= -1. \end{aligned}$$

We define the *multiplier* to be the factor 2. Next, we subtract -1 times the second row from the third to get

$$\begin{aligned} x_1 + x_2 + x_3 &= 1 \\ x_2 + 2x_3 &= 1 \\ 3x_3 &= 0. \end{aligned}$$

In this case, the multiplier is -1. The system is now upper triangular and using back substitution, we obtain $x_3 = 0$, $x_2 = 1$, and $x_1 = 0$. Gaussian elimination can be coded in a straightforward way using matrix notation.

44.2.1 Matrix factorization

There is another way to view Gaussian elimination that is useful for the purposes of programming and handling special cases. Namely, Gaussian elimination is equivalent to computing a *factorization* of the coefficient matrix, $A = LU$, where L is a lower triangular and U an upper triangular $n \times n$ matrix. Given such a factorization of A, solving the system $Ax = b$ is straightforward. We first set $y = Ux$, then solve $Ly = b$ by forward substitution and finally solve $Ux = y$ by backward substitution.

To see that Gaussian elimination gives an LU factorization of A, consider the example above. We performed row operations that brought the system into upper triangular form. If we view these operations as row operations on the matrix A, we get the sequence

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

which is an upper triangular matrix. This is the "U" in the LU decomposition.

The matrix L is determined by the observation that the row operations can be performed by multiplying A on the left by a sequence of special matrices called *Gauss transformations*. These are lower triangular matrices that have at most one nonzero entry in the off-diagonal positions and 1s down the diagonal. We show a Gauss transformation in Fig. 44.3. Multiplying A on the left by the matrix in Fig. 44.3 has the effect of adding

$$\begin{pmatrix}
1 & 0 & & & & \cdots & & & 0 \\
0 & 1 & 0 & & & & & & 0 \\
& \ddots & 1 & \ddots & & 0 & & & \\
\vdots & & & & & & & & \vdots \\
& 0 & 0 & 0 & \ddots & 0 & & & \\
& 0 & \alpha_{ij} & 0 & \ddots & 1 & \ddots & & \\
& 0 & 0 & 0 & & & \ddots & & \\
& & & & & & 0 & 1 & 0 \\
0 & & \cdots & & & & & 0 & 1
\end{pmatrix}$$

FIGURE 44.3. A Gauss transformation.

α_{ij} times row j of A to row i of A. Note that the inverse of this matrix is obtained changing α_{ij} to $-\alpha_{ij}$; we will use this below.

To perform the first row operation on A above, we multiply A on the left by

$$L_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix},$$

to get

$$L_1 A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & -1 & -1 \end{pmatrix}.$$

The effect of pre-multiplication by L_1 is to add $-2\times$ row 1 of A to row 3. Note that L_1 is lower triangular and has ones on the diagonal.

Next we multiply $L_1 A$ on the left by

$$L_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix},$$

and get

$$L_2L_1A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{pmatrix} = U.$$

L_2 is also lower triangular with ones on the diagonal. It follows that $A = L_1^{-1}L_2^{-1}U$ or $A = LU$, where

$$L = L_1^{-1}L_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -1 & 1 \end{pmatrix}.$$

It is easy to see that L is also lower triangular with 1's on the diagonal with the multipliers (with sign change) occurring at the corresponding positions. We thus get the factorization

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

Note that the entries in L below the diagonal are exactly the multipliers used to perform Gaussian elimination on A.

A general linear system can be solved in exactly the same fashion by Gaussian elimination using a sequence of Gauss transformations to obtain a factorization $A = LU$.

An LU factorization can be performed *in situ* using the storage space allotted to the matrix A. The fragment of code shown in Fig. 44.4 computes the LU factorization of A, storing U in the upper triangular part of A and storing the entries in L below the diagonal in the part of A below the diagonal. We illustrate the storage of L and U in Fig. 44.5.

```
for k = 1, ..., n-1, do                     (step through rows)
    for j = k+1, ..., n, do                 (eliminate entries
                                               below diagonal entry)
        a_jk = a_jk/a_kk                    (store the entry of L)
        for m = k+1, ..., n, do             (correct entries
                                               down the row)
            a_jm = a_jm - a_jk × a_km       (store the entry of U)
```

FIGURE 44.4. An algorithm to compute the LU factorization of A that stores the entries of L and U in the storage space of A.

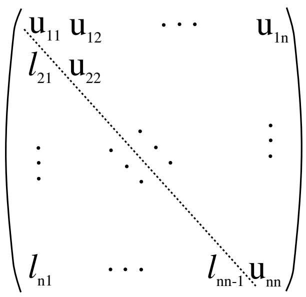


FIGURE 44.5. The matrix output from the algorithm in Fig. 44.4. L and U are stored in the space allotted to A.

44.2.2 *Measuring the cost*

The cost of solving a linear system using a direct method is measured in terms of computer time. In practice, the amount of computer time is proportional to the number of arithmetic and storage operations the computer uses to compute the solution. It is traditional (on a sequential computer) to simplify the cost calculation by equating storing a value, addition, and subtraction and equating multiplication and division when counting operations. Moreover, since multiplication (i.e. multiplications and divisions) generally cost much more than addition on older computers, it is also common to simply count the number of multiplications (=multiplications+divisions).

By this measure, the cost of computing the LU decomposition of an $n \times n$ matrix is $n^3 - n/3 = O(n^3/3)$. We introduce some new notation here, the big "O". The actual count is $n^3/3 - n/3$, however when n is large, the lower order term $-n/3$ becomes less significant. In fact,

$$\lim_{n \to \infty} \frac{n^3/3 - n/3}{n^3/3} = 1, \tag{44.1}$$

and this is the definition of the big "O". (Sometimes the big "O" notation means that the limit of the ratio of the two relevant quantities is any constant). With this notation, the operations count of the LU decomposition is just $O(n^3)$.

The cost of the forward and backward substitutions is much smaller:

44.2.3 *Pivoting*

During Gaussian elimination, it sometimes happens that the coefficient of a variable in the "diagonal position" becomes zero as a result of previous eliminations. When this happens of course, it is not possible to use that equation to eliminate the corresponding entries in the same column lying below the diagonal position. If the matrix is invertible, it is possible to find a non-zero coefficient in the same column and below the diagonal position, and by switching the two rows, the Gaussian elimination can proceed. This is called *zero pivoting*, or just *pivoting*.

Adding pivoting to the LU decomposition algorithm is straightforward. Before beginning the elimination using the current diagonal entry, we check to see if that entry is non-zero. If it is zero, we search the entries below in the same column for the first non-zero value, then interchange the row corresponding to that non-zero entry with the row corresponding to the current diagonal entry which is zero. Because the row interchanges involve rows in the "un-factored" part of A, the form of L and U are not affected. We illustrate this in Fig. 44.6.

```
for k = 1, ..., n-1, do                  (step through rows)
  j=k                                    (search for the first
  while a_jk = 0, j=j+1                    non-zero entry in
                                           the current column)
  for m = 1, ..., n do
    temp = a_km                          (switch the k^th and j^th
    a_km = a_jm                            rows of A)
    a_jm = temp
  for j = k+1, ..., n, do                (eliminate entries
                                           below diagonal entry)
    a_jk = a_jk/a_kk                     (store the entry of L)
    for m = k+1, ..., n, do              (correct entries
                                           down the row)
      a_jm = a_jm - a_jk × a_km          (store the entry of U)
```

FIGURE 44.6. An algorithm to compute the LU factorization of A that used pivoting to avoid zero-valued diagonal entries.

To obtain the correct solution of the linear system $Ax = b$, we have to mirror all pivots performed on A in the data b. This is easy to do with the following trick. We define the vector of integers $p = (1\ 2\ \dots\ n)^\top$. This vector is passed to the LU factorization routine and whenever two rows of A are interchanged, we interchange the corresponding entries in p. After getting the altered p vector back, we pass it to the forward/backward routine. Here, we address the vector b indirectly using the vector p, i.e., we use the vector with entries $(b_{p_i})_{i=1}^n$, which has the effect of interchanging the rows in b in the correct fashion.

There are additional reasons to pivot in practice. As we have noted, the computation of the LU decomposition can be sensitive to errors originating from the finite precision of the computer if the matrix A is close to being non-invertible. We discuss this further below. We mention here however that a special kind of pivoting, called *partial pivoting* can be used to reduce this sensitivity. The strategy behind partial pivoting is to search the entries in the same column and below the current diagonal entry for the largest in absolute value. The row corresponding to the largest entry in magnitude is interchanged with the row corresponding to the current entry at the diagonal. The use of partial pivoting generally gives more accurate results than factorization without partial pivoting. One reason is that partial pivoting insures that the multipliers in the elimination process are kept as small as possible and consequently the errors in each entry are magnified by as little as possible during the course of the Gaussian elimination. We illustrate this with an example. Suppose that we solve

$$\begin{aligned} .000100x_1 + 1.00x_2 &= 1.00 \\ 1.00x_1 + 1.00x_2 &= 2.00 \end{aligned}$$

on a computer that holds three digits. Without pivoting, we get

$$\begin{aligned} .000100x_1 + 1.00x_2 &= 1.00 \\ -10000x_2 &= -10000 \end{aligned}$$

which implies that $x_2 = 1$ and $x_1 = 0$. Note the large multiplier that is required for the elimination. Since the true answer is $x_1 = 1.0001$ and $x_2 = .9999$, the computed result has an error of 100% in x_1. If we switch the two rows before eliminating, which corresponds exactly to the partial pivoting strategy, we get

$$\begin{aligned} 1.00x_1 + 1.00x_2 &= 2.00 \\ 1.00x_2 &= 1.00 \end{aligned}$$

which gives $x_1 = x_2 = 1.00$ as a result.

44.3 Direct methods for special systems

It is often the case that the matrices arising from the Galerkin finite element method applied to a differential equation have special properties that can be useful during the solution of the associated algebraic equations. For example, the stiffness matrix for the Galerkin finite element approximation of the two-point boundary value problem with no convection is symmetric, positive-definite, and tridiagonal. In this section, we examine a couple of different classes of problems that occur frequently.

44.3.1 Symmetric, positive-definite systems

As we mentioned, symmetric, positive-definite matrices are often encountered when discretizing differential equations (especially if the spatial part of the differential equation is of the type called elliptic). If A is symmetric and positive-definite, then it can be factored as $A = BB^\top$ where B is a lower triangular matrix with positive diagonal entries. This factorization can be computed from the LU decomposition of A, but there is a *compact method* of factoring A that requires only $O(n^3/6)$ multiplications called *Cholesky's method.*:

$$
\begin{aligned}
&b_{11} = \sqrt{a_{11}} \\
&b_{i1} = \frac{a_{i1}}{b_{11}}, \quad 2 \le i \le n, \\
&\begin{cases} b_{jj} = \left(a_{jj} - \sum_{k=1}^{j-1} b_{jk}^2\right)^{1/2}, \\ b_{ij} = \left(a_{ij} - \sum_{k=1}^{j-1} b_{ik} b_{jk}\right)/b_{jj}, \end{cases} \qquad 2 \le j \le n,\ j+1 \le i \le n
\end{aligned}
$$

This is called a compact method because it is derived by assuming that the factorization exists and then computing the coefficients of B directly from the equations obtained by matching coefficients in $BB^\top = A$. For example, if we compute the coefficient in the first row and column of $BB^\top$ we get b_{11}^2, which therefore must equal a_{11}. It is possible to do this because A is positive-definite and symmetric, which implies among other things that the diagonal entries of A remain positive throughout the factorization process and pivoting is not required when computing an LU decomposition.

Alternatively, the square roots in this formula can be avoided by computing a factorization $A = CDC^\top$ where C is a lower triangular matrix with ones on the diagonal and D is a diagonal matrix with positive diagonal coefficients.

44.3.2 Banded systems

Banded systems are matrices with non-zero coefficients only in some number of diagonals centered around the main diagonal. In other words, $a_{ij} = 0$ for $j \le i - d_l$ and $j \ge i + d_u$, $1 \le i, j \le n$, where d_l is the *lower bandwidth*, d_u is the *upper bandwidth*, and $d = d_u + d_l - 1$ is called the *bandwidth*. We illustrate this in Fig. 44.7. The stiffness matrix computed for the two-point boundary value problem with no convection is an example of a tridiagonal matrix, which is a matrix with lower bandwidth 2, upper bandwidth 2, and bandwidth 3.

When performing the Gaussian elimination used to compute the LU decomposition, we see that the entries of A that are already zero do not have to be reduced further. If there are only relatively few diagonals with non-zero entries, then the potential saving is great. Moreover, there is no need to store the zero-valued entries of A. It is straightforward to adapt the LU factorization and forward/backward substitution routines to a banded pattern, once a storage scheme has been devised. For example, we can store

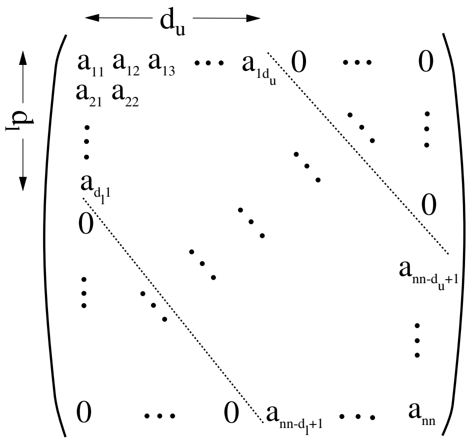


FIGURE 44.7. The notation for a banded matrix.

a tridiagonal matrix as a $3 \times n$ matrix:

$$\begin{pmatrix} a_{21} & a_{31} & 0 & & \cdots & & 0 \\ a_{12} & a_{22} & a_{32} & 0 & \cdots & & 0 \\ 0 & a_{13} & a_{23} & a_{33} & 0 & \cdots & \vdots \\ & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & & 0 & a_{1n-1} & a_{2n-1} & a_{3n-1} \\ 0 & & \cdots & & 0 & a_{1n} & a_{2n} \end{pmatrix}.$$

The routine displayed in Fig. 44.8 computes the LU factorization, while the routine in Fig. 44.9 performs the forward/backward substitution.

```
for k = 2, ..., n, do
    a_1k = a_1k/a_2k-1
    a_2k = a_2k - a_1k × a_3k-1
```

FIGURE 44.8. A routine for computing the LU factorization of a tridiagonal system.

The cost of this routine grows linearly with the dimension, rather than at a cubic rate as in the full case. Moreover, we use only the equivalent of six vectors of dimension n for storage. A more efficient version, derived as a compact method, uses even less.

This algorithm assumes that no pivoting is required to factor A. Pivoting during the factorization of a banded matrix raises the difficulty that the

```
y_1 = b_1
for k = 2, ..., n, do
    y_k = b_k - a_1k × y_k-1

x_n = y_n/a_2n
for k = n-1, ..., 1, do
    x_k = (y_k - a_3k × x_k+1)/a_2k
```

FIGURE 44.9. Using forward and backward substitution to solve a tridiagonal system given the LU factorization.

bandwidth becomes larger. This is easy to see in a tridiagonal matrix, in which case we have to store an extra vector to hold the extra elements above the diagonal that result if two adjacent rows are switched.

As for a tridiagonal matrix, it is straightforward to program special LU factorization and forward/backward substitution routines for a matrix with bandwidth d. The operations count is $O(nd^2/2)$ and the storage requirement is a matrix of dimension $d \times n$ if no pivoting is required. If d is much less than n, the savings in a special approach are considerable.

While it is true that if A is banded, then L and U are also banded, it is also true that in general L and U have non-zero entries in positions where A is zero. This is called *fill-in*. In particular, the stiffness matrix for a boundary value problem in several variables is banded and moreover most of the sub-diagonals in the band have zero coefficients. However, L and U do not have this property and we may as well treat A as if all the diagonals in the band have non-zero entries.

Banded matrices are one example of the class of *sparse* matrices. Recall that a sparse matrix is a matrix with mostly zero entries. As for banded matrices, it is possible to take advantage of sparsity to reduce the cost of factoring A in terms of time and storage. However, it is more difficult to do this than for banded matrices if the sparsity pattern puts non-zero entries at any location in the matrix. One approach to this problem is based on rearranging the equations and variables, or equivalently rearranging the rows and columns to form a banded system.

44.4 Iterative methods

Instead of solving $Ax = b$ directly, we now consider iterative solution methods based on computing a sequence of approximations $x^{(k)}$, $k = 1, 2, ...,$ such that

$$\lim_{k\to\infty} x^{(k)} = x \quad \text{or} \quad \lim_{k\to\infty} \|x^{(k)} - x\| = 0,$$

for some norm $\|\cdot\|$.

Note that the finite precision of a computer has a different effect on an iterative method than it has on a direct method. A theoretically convergent sequence can not reach its limit in general on a computer using a finite number of digits. In fact, at the point at which the change from one iterate to the next occurs outside the range of digits held by the computer, the sequence simply stops changing. Practically speaking, there is no point computing iterations past this point, even if the limit has not been reached. On the other hand, it is often sufficient to have less accuracy than the limit of machine precision, and thus it is important to be able to estimate the accuracy of the current iterate.

44.4.1 Minimization algorithms

We first construct iterative methods for a linear system $Ax = b$ where A is symmetric and positive-definite. In this case, the solution x can be characterized equivalently as the solution of the quadratic minimization problem: find $x \in \mathbb{R}^n$ such that

$$F(x) \leq F(y) \quad \text{for all } y \in R^n, \tag{44.2}$$

where

$$F(y) = \frac{1}{2}(Ay, y) - (b, y),$$

with $(\cdot,\cdot)$ denoting the usual Euclidean scalar product.

We construct an iterative method for the solution of the minimization problem (44.2) based on the following simple idea: given an approximation $x^{(k)}$, compute a new approximation $x^{(k+1)}$ such that $F(x^{(k+1)}) < F(x^{(k)})$. On one hand, since F is a quadratic function, there must be a "downhill" direction from the current position, unless we are at the minimum. On the other hand, we hope that computing the iterates so that their function values are strictly decreasing, will force the sequence to converge to the minimum point x. Such an iterative method is called a *minimization method.*

Writing $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$, where $d^{(k)}$ is a *search direction* and α_k is a *step length*, by direct computation we get

$$F(x^{(k+1)}) = F(x^{(k)}) + \alpha_k \big(Ax^{(k)} - b, d^{(k)}\big) + \frac{\alpha_k^2}{2}\big(Ad^{(k)}, d^{(k)}\big),$$

where we used the symmetry of A to write $(Ax^{(k)}, d^{(k)})=(x^{(k)}, Ad^{(k)})$. If the step length is so small that the second order term in α_k can be neglected, then the direction $d^{(k)}$ in which F decreases most rapidly, or the direction of *steepest descent*, is

$$d^{(k)} = -(Ax^{(k)} - b) = -r^{(k)},$$

which is the opposite direction to the residual error $r^{(k)} = Ax^{(k)} - b$. This suggests using an iterative method of the form

$$x^{(k+1)} = x^{(k)} - \alpha_k r^{(k)}. \tag{44.3}$$

A minimization method with this choice of search direction is called a *steepest descent method.* The direction of steepest descent is perpendicular to the *level curve* of F through $x^{(k)}$, which is the curve in the graph of F generated by the points where F has the same value as at $x^{(k)}$. We illustrate this in Fig. 44.10.

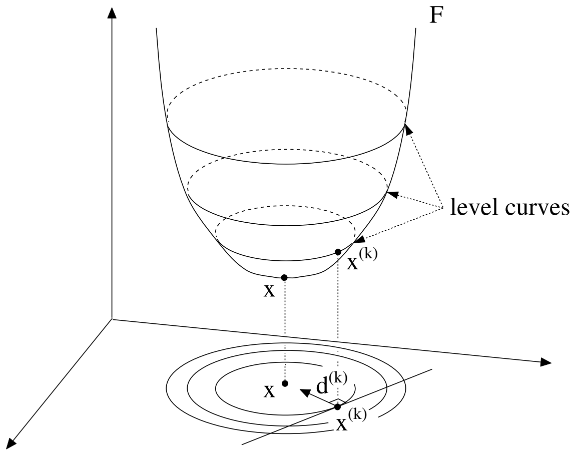


FIGURE 44.10. The direction of steepest descent of F at a point is perpendicular to the level curve of F through the point.

It remains to choose the step lengths α_k. Staying with the underlying principle, we choose α_k to give the minimum value of F in the direction of $d^{(k)}$ starting from $x^{(k)}$. Differentiating $F(x^{(k)} + \alpha_k r^{(k)})$ with respect to α_k and setting the derivative zero gives

$$\alpha_k = -\frac{\big(r^{(k)}, d^{(k)}\big)}{\big(d^{(k)}, Ad^{(k)}\big)}. \tag{44.4}$$

As a simple illustration, we consider the case

$$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \quad 0 < \lambda_1 < \lambda_2, \tag{44.5}$$

and $b = 0$, corresponding to the minimization problem

$$\min_{y \in \mathbb{R}^n} \frac{1}{2}\big(\lambda_1 y_1^2 + \lambda_2 y_2^2\big),$$

with solution $x = 0$.

Applying (44.3) to this problem, we iterate according to

$$x^{(k+1)} = x^{(k)} - \alpha_k A x^{(k)},$$

using for simplicity a constant step length with $\alpha_k = \alpha$ instead of (44.4). In Fig. 44.11, we plot the iterations computed with $\lambda_1 = 1$, $\lambda_2 = 9$, and $x^{(0)} = (9,1)^\top$. The convergence in this case is quite slow. The reason is that if $\lambda_2 \gg \lambda_1$, then the search direction $-(\lambda_1 x_1^{(k)}, \lambda_2 x_2^{(k)})^\top$ and the direction $-(x_1^{(k)}, x_2^{(k)})^\top$ to the solution at the origin, are very different. As a result the iterates swing back and forth across the long, narrow "valley".

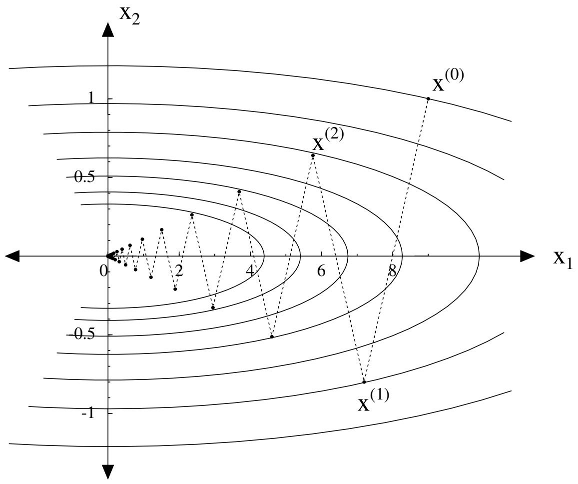


FIGURE 44.11. A sequence generated by the steepest descent method for (44.5) plotted together with some level curves of F.

It turns out that the rate at which the steepest descent method converges in general depends on the *condition number* $\kappa(A) = \lambda_n/\lambda_1$ of A, where $\lambda_1 \le \lambda_2 \le ... \le \lambda_n$ are the eigenvalues of A (counted with multiplicity). In other words, the condition number of a symmetric positive definite matrix is the ratio of the largest eigenvalue to the smallest eigenvalue.

The general definition of the condition number of a matrix A in terms of a norm $\|\cdot\|$ is $\kappa(A) = \|A\|\|A^{-1}\|$. In the $\|\cdot\|_2$ norm, the two definitions are equivalent on symmetric matrices. Using any definition, a matrix is said to be *ill-conditioned* if the $\log(\kappa(A))$ is of the order of the number of digits used in the computer. As we said, we can expect to have difficulty solving an ill-conditioned system; which in terms of direct methods means large errors due to rounding errors and in terms of iterative methods means slow convergence.

We now analyze the steepest descent method for $Ax = b$ in the case of a constant step length α, where we iterate according to

$$x^{(k+1)} = x^{(k+1)} - \alpha(Ax^{(k)} - b).$$

Since the exact solution x satisfies $x = x - \alpha(Ax - b)$, we get the following equation for the error $e^{(k)} = x - x^{(k)}$:

$$e^{(k+1)} = (I - \alpha A)e^{(k)}.$$

The iterative method converges if the error tend to zero. Taking norms, we get

$$\|e^{(k+1)}\| \leq \mu \, \|e^{(k)}\| \tag{44.6}$$

where we use the spectral estimate (**??**) to write

$$\mu = \|I - \alpha A\| = \max_j |1 - \alpha\lambda_j|,$$

since the eigenvalues of the matrix $I - \alpha A$ are $1 - \alpha\lambda_j$, $j = 1, ..., n$. Iterating this estimate we get

$$\|e^{(k+1)}\| \leq \mu^k \, \|e^{(0)}\|, \tag{44.7}$$

where $e^{(0)}$ is the initial error.

To understand when (44.6), or (44.7), guarantees convergence, consider the scalar sequence $\{\lambda^k\}$ for $k \geq 0$. If $|\lambda| < 1$, then $\lambda^k \to 0$; if $\lambda = 1$, then the sequence is always 1; if $\lambda = -1$, the sequence alternates between 1 and -1 and does not converge; and if $|\lambda| > 1$, then the sequence diverges. Therefore if we want the iteration to converge for any initial value, then we must choose α so that $\mu < 1$. Since the λ_j are positive by assumption, $1 - \alpha\lambda_j < 1$ automatically, and we can guarantee that $1 - \alpha\lambda_j > -1$ if α satisfies $\alpha < 2/\lambda_n$. Choosing $\alpha = 1/\lambda_n$, which is not so far from optimal, we get

$$\mu = 1 - 1/\kappa(A).$$

If $\kappa(A)$ is large, then the convergence can be slow because then the reduction factor $1 - 1/\kappa(A)$ is close to one. More precisely, the number of steps required to lower the error by a given amount is proportional to the condition number.

When an iteration converges in this fashion, i.e. the error decreases (more or less) by a given factor in each iteration, then we say that the iteration converges *linearly.* We define the *rate of convergence* to be $-\log(\mu)$. The motivation is that the number of iterations are required to reduce the error by a factor of 10^{-m} is approximately $-m\log(\mu)$. Note that a faster rate of convergence means a smaller value of μ.

This is an *a priori* estimate of the error reduction per iteration, since we estimate the error before the computation. Such an analysis must account for the slowest possible rate of convergence because it holds for all initial vectors.

Consider the system $Ax = 0$ with

$$A = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix}, \tag{44.8}$$

where $0 < \lambda_1 < \lambda_2 < \lambda_3$. For an initial guess $x^{(0)}=(x_1^0, x_2^0, x_3^0)^\top$, the steepest descent method with $\alpha = 1/\lambda_3$ gives the sequence

$$x^{(k)} = \left(\big(1 - \frac{\lambda_1}{\lambda_3}\big)^k x_1^0, \big(1 - \frac{\lambda_2}{\lambda_3}\big)^k x_2^0, 0 \right), \quad k = 1, 2, ...,$$

and,

$$\|e^{(k)}\| = \sqrt{\left(1 - \frac{\lambda_1}{\lambda_3}\right)^{2k} \left(x_1^0\right)^2 + \left(1 - \frac{\lambda_2}{\lambda_3}\right)^{2k} \left(x_2^0\right)^2}, \quad k = 1, 2, ...$$

Thus for a general initial guess, the size of the error is given by the root mean square average of the corresponding iterate and the rate that the errors decrease is the root mean square average of the rates of decrease of the components. Therefore, depending on the initial vector, initially the iterates will generally converge more quickly than the rate of decrease of the first, i.e. slowest, component. In other words, more quickly than the rate predicted by (44.6), which bounds the rate of decrease of the errors by the rate of decrease in the slowest component. However, as the iteration proceeds, the second component eventually becomes much smaller than the first component (as long as $x_1^0 \neq 0$) and we can neglect that term in the expression for the error, i.e.

$$\|e^{(k)}\| \approx \big(1 - \frac{\lambda_1}{\lambda_3}\big)^k |x_1^0| \quad \text{for } k \text{ sufficiently large.} \tag{44.9}$$

In other words, the rate of convergence of the error for almost all initial vectors eventually becomes dominated by the rate of convergence of the slowest component. It is straightforward to show that the number of iterations that we have to wait for this approximation to be valid is determined by the relative sizes of the first and second components of $x^{(0)}$.

This simple error analysis does not apply to the unmodified steepest descent method with varying α_k. However, it is generally true that the rate of convergence depends on the condition number of A, with a larger condition number meaning slower convergence. If we again consider the 2×2 example (44.5) with $\lambda_1 = 1$ and $\lambda_2 = 9$, then the estimate (44.6) for the simplified method suggests that the error should decrease by a factor of $1 - \lambda_1/\lambda_2 \approx .89$ in each iteration. The sequence generated by $x^{(0)} = (9, 1)^\top$ decreases by exactly .8 in each iteration. The simplified analysis overpredicts the rate of convergence for this particular sequence, though not by a lot. By way of comparison, if we choose $x^{(0)} = (1,1)^\top$, we find that the ratio of successive iterations alternates between $\approx .126$ and $\approx .628$, because α_k oscillates in value, and the sequence converges much more quickly than predicted. On the other hand, there are initial guesses leading to sequences that converge at the predicted rate.

The stiffness matrix A of a linear second order two-point boundary value problem with no convection is symmetric and positive-definite, and its condition number $\kappa(A) \propto h^{-2}$. Therefore the convergence of the steepest descent method is very slow if the number of mesh points is large.

44.4.2 A general framework for iterative methods

We now briefly discuss iterative methods for a general, linear system $Ax = b$, following the classical presentation of iterative methods in Isaacson and Keller ([**?**]). Recall that some matrices, like diagonal and triangular matrices, are relatively easy and cheap to invert, and Gaussian elimination can be viewed as a method of factoring A into such matrices. One way to view an iterative method is an attempt to approximate A^{-1} by the inverse of a part of A that is easier to invert. This is called an approximate inverse of A, and we use this to produce an approximate solution to the linear system. Since we don't invert the matrix A, we try to improve the approximate solution by repeating the partial inversion over and over. With this viewpoint, we start by *splitting* A into two parts:

$$A = N - P,$$

where the part N is chosen so that the system $Ny = c$ for some given c is relatively inexpensive to solve. Noting that the true solution x satisfies $Nx = Px + b$, we compute $x^{(k+1)}$ from $x^{(k)}$ by solving

$$Nx^{(k+1)} = Px^{(k)} + b \quad \text{for } k = 1, 2, ..., \tag{44.10}$$

where $x^{(0)}$ is an initial guess. For example, we may choose N to be the diagonal of A:

$$N_{ij} = \begin{cases} a_{ij}, & i = j, \\ 0, & i \neq j, \end{cases}$$

or triangular:

$$N_{ij} = \begin{cases} a_{ij}, & i \geq j, \\ 0, & i < j. \end{cases}$$

In both cases, solving the system $Nx^{(k+1)} = Px^{(k)} + b$ is cheap compared to doing a complete Gaussian elimination on A. so we could afford to do it many times.

As an example, suppose that

$$A = \begin{pmatrix} 4 & 1 & 0 \\ 2 & 5 & 1 \\ -1 & 2 & 4 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}, \tag{44.11}$$

and we choose

$$N = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 4 \end{pmatrix} \quad \text{and} \quad P = \begin{pmatrix} 0 & -1 & 0 \\ -2 & 0 & -1 \\ 1 & -2 & 0 \end{pmatrix},$$

in which case the equation $Nx^{(k+1)} = Px^{(k)} + b$ reads

$$\begin{aligned} 4x_1^{k+1} &= -x_2^k + 1 \\ 5x_2^{k+1} &= -2x_1^k - x_3^k \\ 4x_3^{k+1} &= x_1^k - 2x_2^k + 3. \end{aligned}$$

Being a diagonal system it is easily solved, and choosing an initial guess and computing, we get

$$x^{(0)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \; x^{(1)} = \begin{pmatrix} 0 \\ -.6 \\ .5 \end{pmatrix}, \; x^{(2)} = \begin{pmatrix} .4 \\ -.1 \\ 1.05 \end{pmatrix}, \; x^{(3)} = \begin{pmatrix} .275 \\ -.37 \\ .9 \end{pmatrix},$$

$$x^{(4)} = \begin{pmatrix} .3425 \\ -.29 \\ 1.00375 \end{pmatrix}, \; \cdots \; x^{(15)} = \begin{pmatrix} .333330098 \\ -.333330695 \\ .999992952 \end{pmatrix}, \; \cdots$$

The iteration appears to converge to the true solution $(1/3, -1/3, 1)^\top$.

In general, we could choose $N = N_k$ and $P = P_k$ to vary with each iteration.

To analyze the convergence of (44.10), we subtract (44.10) from the equation $Nx = Px + b$ satisfied by the true solution to get an equation for the error $e^{(k)} = x - x^{(k)}$:

$$e^{(k+1)} = Me^{(k)},$$

where $M = N^{-1}P$ is the *iteration matrix*. Iterating on k gives

$$e^{(k+1)} = M^{k+1}e^{(0)}. \tag{44.12}$$

Rephrasing the question of convergence, we are interested in whether $e^{(k)} \to 0$ as $k \to \infty$. By analogy to the scalar case discussed above, if M is "small", then the errors $e^{(k)}$ should tend to zero. Note that the issue of convergence is independent of the data b.

If $e^{(0)}$ happens to be an eigenvector of M, then it follows from (44.12)

$$\|e^{(k+1)}\| = |\lambda|^{k+1}\|e^{(0)}\|,$$

and we conclude that if the method converges then we must have $|\lambda| < 1$ (or $\lambda = 1$). Conversely, one can show that if $|\lambda| < 1$ for all eigenvalues of M, then the method (44.10) indeed does converge:

Theorem 44.1 *An iterative method converges for all initial vectors if and only if every eigenvalue of the associated iteration matrix is less than one in magnitude.*

This theorem is often expressed using the *spectral radius* $\rho(M)$ of M, which is the maximum of the magnitudes of the eigenvalues of A. An iterative method converges for all initial vectors if and only if $\rho(M) < 1$. In general, the asymptotic limit of the ratio of successive errors computed in $\| \|_\infty$ is close to $\rho(M)$ as the number of iterations goes to infinity. We define the *rate of convergence* to be $R_M = -\log(\rho(M))$. The number of iterations required to reduce the error by a factor of 10^m is approximately m/R_M.

Practically speaking, "asymptotic" means that the ratio can vary as the iteration proceeds, especially in the beginning. In previous examples, we saw that this kind of a priori error result can underestimate the rate of convergence even in the special case when the matrix is symmetric and positive-definite (and therefore has an orthonormal basis of eigenvectors) and the iterative method uses the steepest descent direction. The general case now considered is more complicated, because interactions may occur in direction as well as magnitude, and a spectral radius estimate may overestimate the rate of convergence initially. As an example, consider the non-symmetric (even non-normal) matrix

$$A = \begin{pmatrix} 2 & -100 \\ 0 & 4 \end{pmatrix} \tag{44.13}$$

choosing

$$N = \begin{pmatrix} 10 & 0 \\ 0 & 10 \end{pmatrix} \text{ and } P = \begin{pmatrix} 8 & 100 \\ 0 & 6 \end{pmatrix} \text{ gives } M = \begin{pmatrix} .9 & 10 \\ 0 & .8 \end{pmatrix}.$$

In this case, $\rho(M) = .9$ and we expect the iteration to converge. Indeed it does converge, but the errors become quite large before they start to approach zero. We plot the iterations starting from $x^{(0)} = (1,1)^\top$ in Fig. 44.12.

The goal is obviously to choose an iterative method so that the spectral radius of the iteration matrix is small. Unfortunately, computing $\rho(M)$ in the general case is much more expensive than solving the original linear system and is impractical in general. We recall that $|\lambda| \leq \|A\|$ holds for any norm and any eigenvalue λ of A. The following theorem indicates a practical way to check for convergence.

Theorem 44.2 *Assume that $\|N^{-1}P\| \leq \mu$ for some constant $\mu < 1$ and matrix norm $\| \cdot \|$. Then the iteration converges and $\|e^{(k)}\| \leq \mu^k \|e^{(0)}\|$ for $k \geq 0$.*

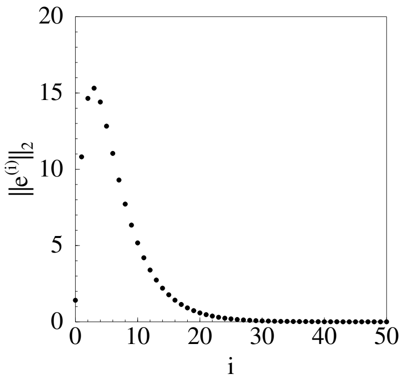


FIGURE 44.12. The results of an iterative method computed using a non-normal matrix.

This theorem is also an a priori convergence result and suffers from the same deficiency as the analysis of the simplified steepest descent method presented above. In fact, choosing an easily computable matrix norm, like $\| \|_\infty$, generally leads to an even more inaccurate estimate of the convergence rate than would be obtained by using the spectral radius. In the worst case, it is entirely possible that $\rho(M) < 1 < \|M\|$ for the chosen norm, and hence the iterative method converges even though the theorem does not apply. The amount of "slack" in the bound in Theorem 44.2 depends on how much larger $\|A\|_\infty$ is than $\rho(A)$.

For the 3×3 example (44.11), we compute $\|N^{-1}P\|_\infty = 3/4 = \lambda$ and therefore we know the sequence converges. The theorem predicts that the error will get reduced by a factor of $3/4$ every iteration. If we examine the error of each iterate along with the ratios of successive errors after the first iteration:

i	$\|e^{(i)}\|_\infty$	$\|e^{(i)}\|_\infty / \|e^{(i-1)}\|_\infty$
0	1.333	
1	.5	.375
2	.233	.467
3	.1	.429
4	.0433	.433
5	.0194	.447
6	.00821	.424
7	.00383	.466
8	.00159	.414
9	.000772	.487

we find that after the first few iterations, the errors get reduced by a factor in the range of .4–.5 each iteration and not the factor $3/4$ predicted above. The ratio of $e^{(40)}/e^{(39)}$ is approximately .469. If we compute the eigenvalues of M, we find that $\rho(M) \approx .476$ which is close to the ratio of successive errors. To decrease the initial error by a factor of 10^{-4} using the predicted decrease of .75 per iteration, we would compute 33 iterations, while only 13 iterations are actually needed.

We get different methods, and different rates of convergence, by choosing different N and P. The method used in the example above is called the *Jacobi* method. In general, this consists of choosing N to be the "diagonal part" of A and P to be the negative of the "off-diagonal" part of A. This gives the set of equations

$$x_i^{k+1} = -\frac{1}{a_{ii}}\Big(\sum_{j\neq i} a_{ij}x_j^k - b_i\Big), \quad i = 1, ..., n.$$

To derive a more sophisticated method, we write out these equations in Fig. 44.13. The idea behind the *Gauss-Seidel* method is to use the new

$$x_1^{k+1} = -\frac{1}{a_{11}}\left(0 + a_{12}x_2^k + \cdots + a_{1n}x_n^k - b_1\right)$$
$$x_2^{k+1} = -\frac{1}{a_{22}}\left(a_{21}\boxed{x_1^k} + 0 + a_{23}x_3^k + \cdots + a_{2n}x_n^k - b_2\right)$$
$$x_3^{k+1} = -\frac{1}{a_{33}}\left(a_{31}\boxed{x_1^k} + a_{32}\boxed{x_2^k} + 0 + a_{34}x_4^k + \cdots - b_3\right)$$
$$\vdots$$
$$x_n^{k+1} = -\frac{1}{a_{nn}}\left(a_{n1}\boxed{x_1^k} + a_{n2}\boxed{x_2^k} + \cdots + a_{nn-1}\boxed{x_{n-1}^k} + 0 - b_n\right)$$

FIGURE 44.13. The Gauss-Seidel method substitutes new values of the iteration as they become available.

values of the approximation in these equations as they become known. The substitutions are drawn in Fig. 44.13. Presumably, the new values are more accurate than the old values, hence we might guess that this iteration will converge more quickly. The equations can be written

$$x_i^{k+1} = \frac{1}{a_{ii}}\Big(-\sum_{j=1}^{i-1} a_{ij}x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij}x_j^k + b_i\Big).$$

If we decompose A into the sum of its lower triangular L, diagonal D, and upper triangular U parts, $A = L+D+U$, then the equations can be written $Dx^{(k+1)} = -Lx^{(k+1)} - Ux^{(k)} + b$ or

$$(D + L)x^{(k+1)} = -Ux^{(k)} + b.$$

Therefore, $N = D + L$ and $P = -U$. The iteration matrix is $M_{\mathrm{GS}} = N^{-1}P = -(D + L)^{-1}U$.

A diagonally dominant matrix often occurs when a parabolic problem is discretized. We have already seen the other case, if A is symmetric and positive-definite then the Gauss-Seidel method converges. This is quite hard to prove, see Isaacson and Keller ([?]) for details.

44.5 Estimating the error of the solution

The issue of estimating the error of the numerical solution of a linear system $Ax = b$ arises both in Gaussian elimination, because of the cumulative effects of round-off errors, and when using iterative methods, where we need a stopping criterion. Therefore it is important to be able to estimate the error in some norm with a fair degree of accuracy.

We discussed this problem in the context of iterative methods in the last section when we analyzed the convergence of iterative methods and Theorem 44.2 gives an *a priori* estimate for the convergence rate. It is an a priori estimate because the error is bounded before the computation begins. Unfortunately, as we saw, the estimate may not be very accurate on a particular computation, and it also requires the size of the initial error. In this section, we describe a technique of *a posteriori* error estimation that uses the approximation after it is computed to give an estimate of the error of that particular approximation.

We assume that x_c is a numerical solution of the system $Ax = b$ with exact solution x, and we want to estimate the error $\|x - x_c\|$ in some norm $\|\cdot\|$. We should point out that we are actually comparing the approximate solution $\tilde{x}_c$ of $\tilde{A}\tilde{x} = \tilde{b}$ to the true solution $\tilde{x}$, where $\tilde{A}$ and $\tilde{b}$ are the finite precision computer representations of the true A and b respectively. The best we can hope to do is compute $\tilde{x}$ accurately. To construct a complete picture, it would be necessary to examine the effects of small errors in A and b on the solution x. To simplify things, we ignore this part of the analysis and drop the $\tilde{}$. In a typical use of an iterative method, this turns out to be reasonable. It is apparently less reasonable in the analysis of a direct method, since the errors arising in direct methods are due to the finite precision. However, the initial error caused by storing A and b on a computer with a finite number of digits occurs only once, while the errors in the arithmetic operations involved in Gaussian elimination occur many times, so even in that case it is not an unreasonable simplification.

We start by considering the *residual error*

$$r = Ax_c - b,$$

which measures how well x_c solves the exact equation. Of course, the residual error of the exact solution x is zero but the residual error of x_c is not zero unless $x_c = x$ by some miracle. We now seek to estimate the unknown error $e = x - x_c$ in terms of the computable residual error r.

By subtracting $Ax - b = 0$ from $Ax_c - b = r$, we get an equation relating the error to the residual error:

$$Ae = -r. \tag{44.14}$$

This is an equation of the same from as the original equation and by solving it numerically by the same method used to compute x_c, we get an approximation of the error e. This simple idea will be used in a more sophisticated form below in the context of a posteriori error estimates for Galerkin methods.

We now illustrate this technique on the linear system arising in the Galerkin finite element discretization of a two-point boundary value problem with no convection. We generate a problem with a known solution so that we can compute the error and test the accuracy of the error estimate. We choose the true solution vector x with components $x_i = \sin(\pi i h)$, where $h = 1/(M+1)$, corresponding to the function $\sin(\pi x)$ and then compute the data by $b = Ax$, where A is the stiffness matrix. We use the Jacobi method, suitably modified to take advantage of the fact that A is tridiagonal, to solve the linear system. We use $\| \, \| = \| \, \|_2$ to measure the error.

We compute the Jacobi iteration until the residual error becomes smaller than a given *residual tolerance* RESTOL. In other words, we compute the residual $r^{(k)} = Ax^{(k)} - b$ after each iteration and stop the process when $\|r^{(k)}\| \leq$ RESTOL. We present computations using the stiffness matrix generated by a uniform discretization with $M = 50$ elements yielding a finite element approximation with an error of .0056 in the l_2 norm. We choose the value of RESTOL so that the error in the computation of the coefficients of the finite element approximation is about 1% of the error of the approximation itself. This is reasonable since computing the coefficients of the approximation more accurately would not significantly increase the overall accuracy of the approximation. After the computation of $x^{(k)}$ is complete, we use the Jacobi method to approximate the solution of (44.14) and compute the estimate of the error.

Using the initial vector $x^{(0)}$ with all entries equal to one, we compute 6063 Jacobi iterations to achieve $\|r\| <$ RESTOL $= .0005$. The actual error of $x^{(6063)}$, computed using the exact solution, is approximately .0000506233. We solve (44.14) using the Jacobi method for 6063 iterations, reporting the value of the error estimate every 400 iterations:

iteration	estimate of error
1	0.00049862
401	0.00026027
801	0.00014873
1201	0.000096531
1601	0.000072106
2001	0.000060676
2401	0.000055328
2801	0.000052825
3201	0.000051653
3601	0.000051105
4001	0.000050849
4401	0.000050729
4801	0.000050673
5201	0.000050646
5601	0.000050634
6001	0.000050628

We see that the error estimate is quite accurate after 6001 iterations and sufficiently accurate for most purposes after 2000 iterations. In general, we do not require as much accuracy in the error estimate as we do in the solution of the system, so the estimation of the accuracy of the approximate solution is cheaper than the computation of the solution.

Since we estimate the error of the computed solution of the linear system, we can stop the Jacobi iteration once the error in the coefficients of the finite element approximation is sufficiently small so that we are sure the accuracy of the approximation will not be affected. This is a reasonable strategy given an estimate of the error. If we do not estimate the error, then the best strategy to guarantee that the approximation accuracy is not affected by the solution error is to compute the Jacobi iteration until the residual error is on the order of roughly 10^{-p}, where p is the number of digits that the computer uses. Certainly, there is not much point to computing further Jacobi iterations after this. If we assume that the computations are made in single precision, then $p \approx 8$. It takes a total of 11672 Jacobi iterations to achieve this level of residual error using the same initial guess as above. In fact, estimating the error and computing the coefficients of the approximation to a reasonable level of accuracy costs significantly less than this crude approach.

This approach can also be used to estimate the error of a solution computed by a direct method, provided the effects of finite precision are included. The added difficulty is that in general the residual error of a solution of a linear system computed with a direct method is small, even if the solution is inaccurate. Therefore, care has to be taken when computing the residual error because the possibility that subtractive cancellation makes the calculation of the residual error itself inaccurate. *Subtractive cancellation* is the name for the fact that the difference of two numbers that agree to the first i places has i leading zeroes. If only the first p digits of the numbers are accurate then their difference can have at most $p-i$ accurate significant digits. This can have severe consequences on the accuracy of the residual error if Ax_c and b agree to most of the digits used by the computer. One way to avoid this trouble is to compute the approximation in single precision and the residual in double precision (which means compute the product Ax_c in double precision, then subtract b). The actual solution of (44.14) is relatively cheap since the factorization of A has already been performed and only forward/backward substitution needs to be done.

Chapter 44 Problems

44.1. Using a similar format, write down algorithms to solve a diagonal system and then a lower triangular system using forward substitution. Determine the number of arithmetic operations needed to compute the solution.

44.2. Prove that multiplying a square matrix A on the left by the matrix in Fig. 44.3 has the effect of adding α_{ij} times row j of A to row i of A. Prove that the inverse of the matrix in Fig. 44.3 is obtained changing α_{ij} to $-\alpha_{ij}$

44.3. Show that the product of two Gauss transformations is a lower triangular matrix with ones on the diagonal and the inverse of a Gauss transformation is a Gauss transformation.

44.4. Solve the system

$$\begin{aligned} x_1 - x_2 - 3x_3 &= 3 \\ -x_1 + 2x_2 + 4x_3 &= -5 \\ x_1 + x_2 &= -2 \end{aligned}$$

by computing an LU factorization of the coefficient matrix and using forward/backward substitution.

44.5. On some computers, dividing two numbers is up to ten times more expensive than computing the reciprocal of the denominator and multiplying the result with the numerator. Alter this code to avoid divisions. Note: the reciprocal of the diagonal element a_{kk} has to be computed just once.

44.6. Write some pseudo-code that uses the matrix generated by the code in Fig. 44.4 to solve the linear system $Ax = b$ using forward/backward substitution. Hint: the only missing entries of L are the 1s on the diagonal.

44.7. Show that the cost of a backward substitution using an upper triangular matrix of dimension $n \times n$ is $O(n^2/2)$.

44.8. Determine the cost of multiplying a $n \times n$ matrix with another.

44.9. One way to compute the inverse of a matrix is based on viewing the equation $AA^{-1} = I$ as a set of linear equations for the columns of A^{-1}. If $a^{(j)}$ denotes the j^{th} column of A^{-1}, then it satisfies the linear system

$$Aa^{(j)} = e_j$$

where e_j is the standard basis vector of $\mathbb{R}^n$ with a one in the j^{th} position. Use this idea to write a pseudo-code for computing the inverse of a matrix using LU factorization and forward/backward substitution. Note that it suffices to compute the LU factorization only once. Show that the cost of computing the inverse in this fashion is $O(4n^3/3)$.

44.10. Write a code that computes the Galerkin polynomial approximation for the first example in Section **??** using LU factorization and forward/backward substitution routines.

44.11. Solve the system

$$\begin{aligned} x_1 + x_2 + x_3 &= 2 \\ x_1 + x_2 + 3x_3 &= 5 \\ -x_1 - 2x_3 &= -1. \end{aligned}$$

This requires pivoting.

44.12. Alter the LU decomposition and forward/backward routines to solve a linear system with pivoting.

44.13. Modify the code in Problem 44.12 to use partial pivoting.

44.14. Consider Problem 44.10. Compute the Galerkin approximation for $q = 1, 2, ..., 9$ using LU factorization and forward/backward substitution with and without partial pivoting.

44.15. Count the cost of Cholesky's method.

44.16. Compute the Cholesky factorization of

$$\begin{pmatrix} 4 & 2 & 1 \\ 2 & 3 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

44.17. Compute the Cholesky factorization of the matrix in the system in Problem **??**.

44.18. Show that the operations count for solving a tridiagonal system using the solver described in Fig. 44.9 is $O(5n)$.

44.19. Find an algorithm to solve a tridiagonal system that stores only four vectors of dimension n.

44.20. A factorization of a tridiagonal solver can be derived as a compact method. Assume that A can be factored as

$$A = \begin{pmatrix} \alpha_1 & 0 & & \cdots & 0 \\ \beta_2 & \alpha_2 & 0 & & \vdots \\ 0 & \beta_3 & \alpha_3 & & \\ \vdots & & & \ddots & 0 \\ 0 & \cdots & 0 & \beta_n & \alpha_n \end{pmatrix} \begin{pmatrix} 1 & \gamma_1 & 0 & \cdots & & 0 \\ 0 & 1 & \gamma_2 & 0 & & \\ \vdots & & \ddots & \ddots & & \\ & & & 1 & \gamma_{n-1} & \\ 0 & \cdots & & 0 & 1 & \end{pmatrix}$$

Multiply out the factors and equate the coefficients to get equations for α, β, and γ. Derive some code based on these formulas.

44.21. Write some code to solve the tridiagonal system resulting from the Galerkin finite element discretization of the two-point boundary value problem in Section **??**. Using $M = 50$, compare the time it takes to solve the system with this tridiagonal solver to the time using a full LU decomposition routine.

44.22. Show that the operations count of a banded solver for a $n \times n$ matrix with bandwidth d is $O(nd^2/2)$.

44.23. Write code to solve a linear system with bandwidth five centered around the main diagonal. What is the operations count for your code?

44.24. Prove that the solution of (44.2) is also the solution of $Ax = b$.

44.25. Prove that the direction of steepest descent for a function F at a point is perpendicular to the level curve of F through the same point.

44.26. Prove (44.4).

44.27. Prove that the level curves of F in the case of (44.5) are ellipses with major and minor axes proportional to $1/\sqrt{\lambda_1}$ and $1/\sqrt{\lambda_2}$, respectively.

44.28. Compute the iteration corresponding to $\lambda_1 = 1$, $\lambda_2 = 2$, $\lambda_3 = 3$, and $x^{(0)} = (1,1,1)^\top$ for the system $Ax = 0$ with A defined in (44.8). Make a plot of the ratios of successive errors versus the iteration number. Do the ratios converge to the ratio predicted by the error analysis?

44.29. Prove that the estimate (44.9) generalizes to any symmetric positive-definite matrix A, diagonal or not. Hint: use the fact that there is a set of eigenvectors of A that form an orthonormal basis for $\mathbb{R}^n$ and write the initial vector in terms of this basis. Compute a formula for the iterates and then the error.

44.30. (a) Compute the steepest descent iterations for (44.5) corresponding to $x^{(0)} = (9,1)^\top$ and $x^{(0)} = (1,1)^\top$, and compare the rates of convergence. Try to make a plot like Fig. 44.11 for each. Try to explain the different rates of convergence.

(b) Find an initial guess which produces a sequence that decreases at the rate predicted by the simplified error analysis.

44.31. Prove that the method of steepest descent corresponds to choosing

$$N = N_k = \frac{1}{\alpha_k} I, \text{ and } P = P_k = \frac{1}{\alpha_k} I - A,$$

with suitable α_k in the general iterative solution algorithm.

44.32. Compute the eigenvalues and eigenvectors of the matrix A in (44.13) and show that A is not normal.

44.33. Prove that the matrix $\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$ is normal.

44.34. Prove Theorem 44.2.

44.35. Compute 10 Jacobi iterations using the A and b in (44.11) and the initial guess $x^{(0)} = (-1,\ 1,\ -1)^\top$. Compute the errors and the ratios of successive errors and compare to the results above.

44.36. Repeat Problem 44.35 using

$$A = \begin{pmatrix} 4 & 1 & 100 \\ 2 & 5 & 1 \\ -1 & 2 & 4 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 0 \\ 3 \end{pmatrix}.$$

Does Theorem 44.2 apply to this matrix?

44.37. Show that for the Jacobi iteration, $N = D$ and $P = -(L + U)$ and the iteration matrix is $M_{\mathrm{J}} = -D^{-1}(L + U)$

44.38. (a) Solve (44.11) using the Gauss-Seidel method and compare the convergence with that of the Jacobi method. Also compare $\rho(M)$ for the two methods. (b) Do the same for the system in Problem 44.36.

44.39. (Isaacson and Keller ([?])) Analyze the convergence of the Jacobi and Gauss-Seidel methods for the matrix

$$A = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

in terms of the parameter ρ.

In general it is difficult to compare the convergence of the Jacobi method with that of the Gauss-Seidel method. There are matrices for which the Jacobi method converges and the Gauss-Seidel method fails and vice versa. There are two special classes of matrices for which convergence can be established without further computation. A matrix A is *diagonally dominant* if

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}|, \quad i = 1, ..., n.$$

If A is diagonally dominant then the Jacobi method converges.

44.40. Prove this claim.

44.41. Derive an algorithm that uses the Jacobi method to solve a tridiagonal system. Use as few operations and as little storage as possible.

44.42. Devise an algorithm to estimate the error of the solution of a linear system using single and double precision as suggested. Repeat the example using a tridiagonal solver and your algorithm to estimate the error.

45
Linear Algebra Tool Bag

45.1 Linear algebra in $\mathbb{R}^2$

Scalar product of two vectors $a = (a_1, a_2)$ and $b = (b_1, b_2)$ in $\mathbb{R}^2$:

$$a \cdot b = (a, b) = a_1 b_1 + a_2 b_2.$$

Norm: $|a| = (a_1^2 + a_2^2)^{1/2}$.

Angle between two vectors a and b in $\mathbb{R}^3$: $\cos(\theta) = \frac{a \cdot b}{|a|\,|b|}$.
The vectors a and b are orthogonal if and only if $a \cdot b = 0$.

Vector product of two vectors $a = (a_1, a_2)$ and $b = (b_1, b_2)$ in $\mathbb{R}^3$:

$$a \times b = a_1 b_2 - a_2 b_1.$$

Properties of vector product: $|a \times b| = |a||b||\sin(\theta)|$, where θ is the angle between a and b. In particular, a and b are parallel if and only if $a \times b = 0$.

Volume of parallelogram spanned by two vectors $a, b \in \mathbb{R}^2$:

$$V(a, b) = |a \times b| = |a_1 b_2 - a_2 b_1|.$$

45.2 Linear algebra in $\mathbb{R}^3$

Scalar product of two vectors $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$ in $\mathbb{R}^3$:

$$a \cdot b = \sum_{i=1}^{3} a_i b_i = a_1 b_1 + a_2 b_2 + a_3 b_3.$$

Norm:$|a| = (a_1^2 + a_2^2 + a_3^2)^{1/2}$.

Angle between two vectors a and b in $\mathbb{R}^3$: $\cos(\theta) = \frac{a \cdot b}{|a|\,|b|}$.
The vectors a and b are orthogonal if and only if $a \cdot b = 0$.
Vector product of two vectors $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3)$ in $\mathbb{R}^3$:

$$a \times b = (a_2 b_3 - a_3 b_2, a_3 b_1 - a_1 b_3, a_1 b_2 - a_2 b_1).$$

Properties of vector product: The vector product $a \times b$ of two non-zero vectors a and b in $\mathbb{R}^3$, is orthogonal to both a and b, and $|a \times b| = |a||b||\sin(\theta)|$, where θ is the angle between a and b. In particular, a and b are parallel if and only if $a \times b = 0$.

Volume of parallelepiped spanned by three vectors $a, b, c \in \mathbb{R}^3$:

$$V(a, b, c) = |c \cdot (a \times b)|.$$

45.3 Linear algebra in $\mathbb{R}^n$

Definition of $\mathbb{R}^n$: The set of ordered n-tuples, $x = (x_1, ..., x_n)$ with components $x_i \in \mathbb{R}$, $i = 1, ..., n$.
Vector addition and scalar multiplication: For $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$ in $\mathbb{R}^n$ and $\lambda \in \mathbb{R}$, we define

$$x + y = (x_1 + y_1, x_2 + y_2, ..., x_n + y_n), \quad \lambda x = (\lambda x_1, ..., \lambda x_n).$$

Scalar product:

$$x \cdot y = (x, y) = \sum_{i=1}^{n} x_i y_i.$$

Norm:$|a| = (\sum_{i=1}^{n} a_i^2)^{1/2}$.

Cauchy's inequality:

$$|(x, y)| \leq |x|\,|y|.$$

Angle of two vectors x and y in $\mathbb{R}^n$: $\cos(\theta) = \frac{(x,y)}{|x||y|}$.

Standard basis: $\{e_1, ..., e_n\}$, where $e_i = (0, 0, .., 0, 1, 0, ..., 0)$ with a single coefficient 1 at position i.

Linear independence: A set $\{a_1, ..., a_n\}$ of vectors in $\mathbb{R}^m$ is said to be *linearly independent* if none of the vectors a_i can be expressed as a linear combination of the others, that is, if $\sum_{i=1}^n \lambda_i a_i = 0$ with $\lambda_i \in \mathbb{R}$ implies that $\lambda_i = 0$ for $i = 1, ..., n$.

A basis for $\mathbb{R}^n$ is a linearly independent set of vectors whose linear combinations span $\mathbb{R}^n$. Any basis of $\mathbb{R}^n$ has n elements. Further, a set of n vectors in $\mathbb{R}^n$ span $\mathbb{R}^n$ if and only if it is linearly independent, that is, a set of n vectors in $\mathbb{R}^n$ that spans $\mathbb{R}^n$ or is independent, must be a basis. Also, a set of fewer than n vectors in $\mathbb{R}^n$ cannot span $\mathbb{R}^n$, and a set of more than n vectors in $\mathbb{R}^n$ must be linearly dependent.

45.4 Linear transformations and matrices

An $m \times n$ real (or complex) *matrix* $A = (a_{ij})$ is rectangular array with rows $(a_{i1}, ..., a_{in})$, $i = 1, ..., m$, and columns $(a_{1j}, ..., a_{mj})$, $j = 1, ..., n$, where $a_{ij} \in \mathbb{R}$ (or $a_{ij} \in \mathbb{C}$).

Matrix addition: Given two $m \times n$ matrices $A = (a_{ij})$ and $B = (b_{ij})$, we define $C = A+B$ as the $m \times n$ matrix $C = (c_{ij})$ with elements $c_{ij} = a_{ij}+b_{ij}$, corresponding to elementwise addition.

Multiplication by scalar Given a $m \times n$ matrix $A = (a_{ij})$ and a real number λ, we define the $m \times n$ matrix λA with elements (λa_{ij}), corresponding to multiplying all elements of A by the real number λ.

Matrix multiplication: Given a $m \times p$ matrix A and a $p \times n$ matrix B we define a $m \times n$ matrix AB with elements $(AB)_{ij} = \sum_{k=1}^p a_{ik} b_{kj}$. Matrix multiplication is not commutative, that is, $AB \neq BA$ in general. In particular, BA is defined only if $n = m$.

A linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ can be expressed as $f(x) = Ax$, where $A = (a_{ij})$ is an $m \times n$ matrix with elements $a_{ij} = f_i(e_j) = (e_i, f(e_j))$, where $f(x) = (f_1(x), ..., f_m(x))$. If $g : \mathbb{R}^n \to \mathbb{R}^p$ and $f : \mathbb{R}^p \to \mathbb{R}^m$ are two linear transformations with corresponding matrices A and B, then the matrix of $f \circ g : \mathbb{R}^n \to \mathbb{R}^m$ is given by AB.

Transpose: If $A = (a_{ij})$ is a real $m \times n$ matrix, then the transpose $A^\top$ is an $n \times m$ matrix with elements $a^\top_{ji} = a_{ij}$, and $(Ax, y) = (x, A^\top y)$ for all $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$.

Matrix norms:

$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^m |a_{ij}|, \quad \|A\|_\infty = \max_{i+1,\ldots,m} \sum_{j=1}^n |a_{ij}|, \quad \|A\| = \max_{x \in \mathbb{R}^n} \frac{\|Ax\|}{\|x\|}.$$

If $A = (\lambda_i)$ is a diagonal $n \times n$ matrix with diagonal elements $a_{ii} = \lambda_i$, then

$$\|A\| = \max_{i=1,\ldots,n} |\lambda_i|.$$

Lipschitz constant of a linear transformation: The Lipschitz constant of a linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ given by a $m \times n$ matrix $A = (a_{ij})$ is equal to $\|A\|$.

45.5 Determinant and volume

The determinant det A of an $n \times n$ matrix $A = (a_{ij})$, or the volume $V(a_1, ..., a_n)$ spanned by the column vectors of A, is defined by

$$\det A = V(a_1, ..., a_n) = \sum_\pi \pm a_{\pi(1)\,1} a_{\pi(2)\,2} \cdots a_{\pi(n)\,n},$$

where we sum over all permutations π of the set $\{1, ..., n\}$, and the sign indicates if the permutation is even (+) or odd (-). We have $\det A = \det A^\top$.

Volume $V(a_1, a_2)$ in $\mathbb{R}^2$:

$$\det A = V(a_1, a_2) = a_{11}a_{22} - a_{21}a_{12}.$$

Volume $V(a_1, a_2, a_3)$ in $\mathbb{R}^3$:

$$\begin{aligned}&\det A = V(a_1, a_2, a_3) = a_1 \cdot a_2 \times a_3\\ &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}).\end{aligned}$$

Volume $V(a_1, a_2, a_3, a_4)$ in $\mathbb{R}^4$:

$$\begin{aligned}\det A = V(a_1, a_2, a_3, a_4) &= a_{11}V(\hat{a}_2, \hat{a}_3, \hat{a}_4) - a_{12}V(\hat{a}_1, \hat{a}_3, \hat{a}_4)\\ &+ a_{13}V(\hat{a}_1, \hat{a}_2, \hat{a}_4) - a_{14}V(\hat{a}_1, \hat{a}_2, \hat{a}_3),\end{aligned}$$

where the $\hat{a}_j$, $j = 1, 2, 3, 4$ are the 3-column vectors corresponding to cutting out the first coefficient of the a_j.

Determinant of a triangular matrix: If $A = (a_{ij})$ is a *upper triangular* $n \times n$ matrix, that is $a_{ij} = 0$ for $i > j$, then

$$\det A = a_{11}a_{22} \cdots a_{nn}.$$

This formula also applies to a *lower triangular* $n \times n$ matrix $A = (a_{ij})$ with $a_{ij} = 0$ for $i < j$.

The magic formula: $\det AB = \det A \det B$

Test of linear independence: A set $\{a_1, a_2, ..., a_n\}$ of n vectors in $\mathbb{R}^n$ is linearly independent if and only if $V(a_1, ..., a_n) \neq 0$. The following statements are equivalent for an $n \times n$ matrix A: (a) The columns of A are linearly independent, (b) If $Ax = 0$, then $x = 0$, (c) $\det A \neq 0$.

45.6 Cramer's formula

If A is a $n \times n$ non-singular matrix with $\det A \neq 0$, then the system of equations $Ax = b$ has a unique solution $x = (x_1, ..., x_n)$ for any $b \in \mathbb{R}^n$. given by

$$x_i = \frac{V(a1, ..., a_{i-1}, b, a_{i+1}, ..., a_n)}{V(a_1, a_2, ..., a_n)}, \quad i = 1, ..., n.$$

45.7 Inverse

A nonsingular $n \times n$ matrix A has a inverse matrix A^{-1} satisfying:

$$A^{-1}A = AA^{-1} = I,$$

where I is the $n \times n$ identity matrix.

45.8 Projection

The projection $Pv \in V$ of $v \in \mathbb{R}^n$, where V is a linear subspace of $\mathbb{R}^n$, is uniquely defined defined by $(v - Pv, w) = 0$ for all $w \in V$ and satisfies $|v - Pv| \leq |v - w| \quad$ for all $w \in V$. Further, $PP = P$.

45.9 The Fundamental Theorem of Linear Algebra

If A is a $m \times n$ matrix with null space $N(A) = \{x \in \mathbb{R}^n : Ax = 0\}$ and range $R(A) = \{y = Ax : x \in \mathbb{R}^n\}$, then

$$N(A) \oplus R(A^\top) = \mathbb{R}^n \quad N(A^\top) \oplus R(A) = \mathbb{R}^m,$$

$$\dim\, N(A) + \dim\, R(A^\top) = n, \ \dim\, N(A^\top) + \dim\, R(A) = m,$$

$$\dim\, N(A) + \dim\, R(A) = n, \ \dim\, N(A^\top) + \dim\, R(A^\top) = m,$$

$$\dim\, R(A) = \dim\, R(A^\top),$$

The number of linearly independent columns of A is equal to the number of linearly independent rows of A.

45.10 The QR-decomposition

An $n \times m$ matrix A can be expressed in the form

$$A = QR,$$

where Q is a $n \times m$ matrix with orthogonal columns and R is a $m \times m$ upper triangular matrix.

45.11 Change of basis

A linear transformation $f : \mathbb{R}^n \to \mathbb{R}^n$, with matrix A with respect to the standard basis, has the following matrix in a basis $\{s_1, ...s_n\}$:

$$S^{-1}AS,$$

where the the coefficients s_{ij} of the matrix $S = (s_{ij})$ are the coordinates of the basis vectors s_j with respect to the standard basis.

45.12 Least squares method

The least squares solution of the linear system $Ax = b$ with A an $m \times n$ matrix minimizing $|Ax - b|^2$ satisfies $A^\top Ax = A^\top b$, and is unique if the columns of A are linearly independent.

45.13 Eigenvalues and eigenvectors

If A is an $n \times n$ matrix and $x =\in \mathbb{R}^n$ is a non-zero vector which satisfies

$$Ax = \lambda x,$$

where λ is a real number, then we say that $x \in \mathbb{R}^n$ is an *eigenvector* of A and that λ is a corresponding *eigenvalue* of A. The number λ is an eigenvalue of the $n \times n$ matrix A if and only if λ is a root of the characteristic equation $\det(A - \lambda I) = 0$.

45.14 The Spectral Theorem

If A is a symmetric $n \times n$ matrix A, then there is an orthonormal basis $\{q_1, ..., q_n\}$ of $\mathbb{R}^n$ consisting of eigenvectors q_j of A with corresponding real eigenvalues λ_j, satisfying $Aq_j = \lambda_j q_j$, for $j = 1, ..., n$. We have $D = Q^{-1}AQ$

and $A = QDQ^{-1}$, where Q is the orthogonal matrix with the eigenvectors q_j in the standard basis forming the columns, and D is the diagonal matrix with the eigenvalues λ_j on the diagonal. Further, $\|A\| = \max_{i=1,\ldots,n} |\lambda_i|$.

46
The matrix exponential $\exp(xA)$

> I tell them that if they will occupy themselves with the study of mathematics, they will find in it the best remedy against the lusts of the flesh. (Thomas Mann (1875-1955))

An important special case of the general initial value problem (40.1) is the linear system

$$u'(x) = Au(x) \quad \text{for } 0 < x \leq T,\ u(0) = u_0, \tag{46.1}$$

where A is a *constant* $d \times d$ matrix, $u_0 \in \mathbb{R}^d$, $T > 0$ and the solution $u(x) \in \mathbb{R}^d$ is a column vector. By the general existence result of the previous chapter we know that that a unique solution exists. Recalling that the solution of the scalar problem $u' = au$, a constant, is $u = e^{xa}u_0$, we denote the solution of (46.1) by

$$u(x) = \exp(xA)u_0 = e^{xA}u_0. \tag{46.2}$$

This definition can be extended to $x < 0$ in the obvious way.

To make sense of $\exp(xA)$, we may view $\exp(xA) = e^{xA}$ as the $d \times d$ *matrix* with column i denoted by $\exp(xA)_i$ being the solution vector $u(x)$ with initial data $u_0 = e_i$, where the e_i are the standard basis vectors. This means that $\exp(xA)_i = \exp(xA)e_i$. By linearity we can write the solution $u(x)$ with general initial data $u_0 = \sum_{i=1}^d u_{0,i}e_i$ in the form

$$u(x) = \exp(xA)u_0 = \sum_{i=1}^d \exp(xA)u_{0,i}e_i = \sum_{i=1}^d u_{0,i}\exp(xA)_i.$$

EXAMPLE 46.1. If A is diagonal with diagonal elements d_i, then $exp(xA)$ is a diagonal matrix with diagonal elements $\exp(xd_i)$.

We may express the basic property of the matrix exponential $\exp(xA)$ as follows:

$$\frac{d}{dx}\exp(xA) = A\exp(xA) = Ae^{xA}, \quad \text{for } x \in \mathbb{R} \tag{46.3}$$

We also note the following related basic property, which generalizes a familiar property of the usual exponential (for a proof, see Problem 46.1):

$$\exp(xA)\exp(yA) = \exp((x+y)A) \quad \text{for } x, y \in \mathbb{R}. \tag{46.4}$$

46.1 Analytic computation of $\exp(xA)$ if A is diagonalizable

We have defined $\exp(xA)$ through the solution of the initial value problem (46.1), but we do not yet have an analytical formula for $\exp(xA)$, except in the "trivial" case with A diagonal. It turns out that we can find a formula in the case with A diagonalizable. This formula helps to give an idea of the structure of the matrix function $\exp(xA)$ in terms of the eigenvalues and eigenvectors of A, and in particular gives a chance of identifying cases when $\exp(xA)$ is exponentially decaying as x increases.

We consider the system (46.1) with the matrix A diagonalizable so that there is a nonsingular matrix S such that $S^{-1}AS = D$ or $A = SDS^{-1}$, where D is a diagonal matrix with diagonal elements d_i (the eigenvalues of A), and the columns of S corresponding eigenvectors, see the Chapter The Spectral Theorem. If A is symmetric then S may be chosen to be orthogonal with $S^{-1} = S^\top$. Introducing the new dependent variable $v = S^{-1}u$ so that $u = Sv$, we can rewrite $\dot{u} = Au$ in the form $\dot{v} = S^{-1}ASv = Dv$, and letting $\exp(xD)$ be the diagonal matrix with diagonal elements equal to $\exp(xd_i)$, we have $v(x) = \exp(xD)v(0)$, which we can write in the form $S^{-1}u(x) = \exp(xD)S^{-1}u_0$, that is

$$u(x) = S\exp(xD)S^{-1}u_0.$$

Since according to the previous section we have also decided to write $u(x) = \exp(xA)u_0$, we conclude that

$$\exp(xA) = S\exp(xD)S^{-1}.$$

This shows that indeed $\exp(xA)$ may be viewed as a matrix, and we also get an analytic formula to compute $\exp(xA)$ without having to directly solve $\dot{u} = Ax$. We note that each element of $\exp(xA)$ is a certain linear combination of terms of the form $\exp(xd_i)$ with d_i and eigenvalue of A. We give some basic examples.

EXAMPLE 46.2. (A symmetric with real eigenvalues) Suppose

$$A = \begin{pmatrix} a & 1 \\ 1 & a \end{pmatrix}.$$

The eigenvalues of A are $d_1 = a-1$ and $d_2 = a+1$ and the corresponding matrix S of normalized eigenvectors (which is orthogonal since A is symmetric), is given by

$$S = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad S^{-1} = S^T = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

We compute and find that

$$\exp(xA) = S\exp(xD)S^T = \exp(ax)\begin{pmatrix} \cosh(x) & \sinh(x) \\ \sinh(x) & \cosh(x) \end{pmatrix},$$

and we see that each element of $\exp(xA)$ is a linear combination of $\exp(d_j x)$ with d_j an eigenvalue of A. If $a < -1$, then all elements of $\exp(xA)$ are exponentially decaying.

EXAMPLE 46.3. (A anti-symmetric with purely imaginary eigenvalues)

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

The eigenvalues of A are purely imaginary, $d_1 = -i$ and $d_2 = i$, and the corresponding matrix S of eigenvectors is given by

$$S = \begin{pmatrix} -i & i \\ 1 & 1 \end{pmatrix}, \quad S^{-1} = \frac{1}{2i}\begin{pmatrix} -1 & i \\ 1 & i \end{pmatrix}$$

We compute and find that

$$\exp(xA) = S\exp(xD)S^{-1} = \begin{pmatrix} \cos(x) & -\sin(x) \\ \sin(x) & \cos(x) \end{pmatrix},$$

and we see again that each element of $\exp(xA)$ is a linear combination of $\exp(d_j x)$ with d_j an eigenvalue of A. In this case the problem $\dot{u} = Au$ takes the form of the scalar linear oscillator $\ddot{v} + v = 0$ with $u_1 = v$ and $u_2 = \dot{v}$.

EXAMPLE 46.4. (A non-normal) The matrix

$$A = \begin{pmatrix} a & 1 \\ \epsilon^2 & a \end{pmatrix}$$

with $a \in \mathbb{R}$, and ϵ small positive has the eigenvalues $d_1 = a - \epsilon$ and $d_2 = a + \epsilon$, and the corresponding matrix of eigenvectors S is given by

$$S = \begin{pmatrix} 1 & 1 \\ -\epsilon & \epsilon \end{pmatrix}, \quad S^{-1} = \frac{1}{2}\begin{pmatrix} 1 & \epsilon^{-1} \\ 1 & -\epsilon^{-1} \end{pmatrix}.$$

Computing we find that

$$\exp(xA) = S\exp(xD)S^{-1} = \exp(ax)\begin{pmatrix}\cosh(\epsilon x) & \epsilon^{-1}\sinh(\epsilon x)\\ \epsilon\sinh(\epsilon x & \cosh(\epsilon x)\end{pmatrix}.$$

Again we note that each element of $\exp(xA)$ is a linear combination of $\exp(d_jx)$ with d_j and eigenvalue. We further that if ϵ is small, then S is nearly singular (the two eigenvectors are almost parallel), and the inverse S^{-1} contains large numbers (ϵ^{-1}).

46.2 Properties of $exp(Ax)$

If $A = D = (d_j)$ is diagonal with diagonal elements d_i, then the nature of $\exp(xA)$ is easy to describe. In this case, $\exp(xA)$ is diagonal and if $d_j > 0$, then the corresponding diagonal element $\exp(d_jx)$ is exponentially increasing and if $d_j < 0$, then $\exp(d_jx)$ is exponentially decreasing. If all the d_j are positive (negative), then we may describe $\exp(xA)$ as exponentially increasing (decaying). If $d_i = a+ib$ is complex with $b \neq 0$, then $\exp(d_jx) = \exp(ax)\exp(ibx)$ oscillates with an exponentially increasing or decreasing amplitude depending on the sign of $a > 0$.

If A is diagonalizable with $S^{-1}AS = D$ with $D = (d_j)$ diagonal, then $\exp(xA) = S\exp(xD)S^{-1}$, and it follows that the elements of $\exp(AX)$ are a certain linear combinations of the exponentials $\exp(d_jx)$. If all the d_j are negative, then all elements of $\exp(xA)$ will be exponentially decaying. We will pay particular attention to this case below.

If A is not diagonalizable, then the structure of $\exp(xA)$ is more complex: the elements of $\exp(xA)$ will then be of the form $p(x)\exp(d_jx)$ with d_j an eigenvalue of A and $p(x)$ a polynomial with degree less than the multiplicity of d_j.

46.3 Duhamel's formula

We can generalize the previous discussion to the non-homogeneous problem

$$u'(x) = Au(x) + f(x) \quad \text{for } 0 < x \leq 1,\ u(0) = u_0, \tag{46.5}$$

where $f(x)$ is a given function. The solution $u(x)$ can be expressed in the form of a *Duhamel formula* generalizing the formula that holds for a scalar problem,

$$u(x) = \exp(xA)u_0 + \int_0^x \exp((x-y)A)f(y)\,dy. \tag{46.6}$$

This is readily verified by differentiation,

$$\begin{aligned}\dot{u}(x) &= \frac{d}{dx}\exp(xA)u_0 + \frac{d}{dx}\int_0^x \exp((x-y)A)f(y)\,dy \\ &= A\exp(xA)u_0 + \exp((x-x)A)f(x) + \int_0^x A\exp((x-y)A)f(y)\,dy \\ &= Au(x) + f(x).\end{aligned} \tag{46.7}$$

Chapter 46 Problems

46.1. Prove (46.4). Hint: Assuming that $u' = Au$ we may write $u(x+y) = \exp(xA)u(y) = \exp(xA)\exp(yA)u(0)$ and also $u(x+y) = \exp((x+y)A)u(0)$.

46.2. Rewrite the second order scalar constant coefficient problem $\ddot{v}+a_1\dot{v}+a_0v = 0$ in first order system form $\dot{u} = Au$ by setting $u_1 = v$ and $u_2 = \dot{v}$, and connect the analysis of this chapter to the analysis of the linear oscillator in the Chapter N-body systems. Generalize to higher order scalar problems.

47

Lagrange and the Principle of Least Action*

> Dans les modifications des mouvements, l'action devient ordinairement un Maximum ou un Minimum. (Leibniz)
>
> Whenever any action occurs in nature, the quantity of action employed by this change is the least possible. (Maupertuis 1746)
>
> From my earliest recollection I have had an irresistible liking for mechanics and the physical laws on which mechanics as a science is based. (Reynolds)

47.1 Introduction

Lagrange (1736-1813), see Fig. 47.1, found a way to formulate certain dynamical problems in mechanics using a *Principle of Least Action*. This principle states that the *state* $u(t)$ of a system changes with time t over a given time interval $[t_1, t_2]$, so that the *action integral*

$$I(u) = \int_{t_1}^{t_2} (T(\dot{u}(t)) - V(u(t))\, dt \qquad (47.1)$$

is *stationary*, where $T(\dot{u}(t))$ with $\dot{u} = \frac{du}{dt}$ is the *kinetic energy*, and $V(u(t))$ is the *potential energy* of the *state* $u(t)$. We here assume that the state $u(t)$ is a function $u : [t_1, t_2] \to \mathbb{R}$ satisfying $u(t_1) = u_1$ and $u(t_2) = u_2$, where u_1 and u_2 are given initial and final values. For example, $u(t)$ may be the position of a moving mass at time t. The action integral of a state is thus the difference between the kinetic and potential energies integrated in time along the state.

We shall now get acquainted with Lagrange's famous Principle of Least Action and we shall see that it may be interpreted as a reformulation of Newton's law stating that mass times acceleration equals force. To this end, we first need to explain what is meant by the statement that the *action integral is stationary* for the actual solution $u(t)$. Our tool is Calculus, at its best!

FIGURE 47.1. Lagrange, Inventor of the Principle of Least Action

Following in the foot-steps of Lagrange, consider a *perturbation* $v(t) = u(t)+\epsilon w(t) = (u+\epsilon w)(t)$ of the state $u(t)$, where $w(t)$ is a function on $[t_1, t_2]$ satisfying $w(t_1) = w(t_2) = 0$ and ϵ is a small parameter. The function $v(t)$ corresponds to changing $u(t)$ with the function $\epsilon w(t)$ inside (t_1, t_2) while keeping the values $v(t_1) = u_1$ and $v(t_2) = u_2$. The Principle of Least Action states that the actual path $u(t)$ has the property that for all such functions $w(t)$, we have

$$\frac{d}{d\epsilon} I(u + \epsilon w) = 0 \quad \text{for } \epsilon = 0. \tag{47.2}$$

The derivative $\frac{d}{d\epsilon} I(u + \epsilon w)$ at $\epsilon = 0$, measures the rate of change with respect to ϵ at $\epsilon = 0$ of the value of the action integral with $u(t)$ replaced by $v(t) = u(t) + \epsilon w(t)$. The Principle of Least Action says this rate of change is zero if u is the actual solution, which expresses the stationarity of the action integral.

We now present a couple of basic applications illustrating the use of the Principle of Least Action.

47.2 Mass-spring system

We consider a system of a mass m sliding on a on a horizontal friction-less x-axis and being connected to the origin with a weight-less Hookean spring with spring constant k, see the Chapter Galileo, Newton et al. We know that this system may be described by the equation $m\ddot{u} + ku = 0$, where $u(t)$ is the length of the spring at time t. We derive this model by using the Principle of Least Action. In this case,

$$T(\dot{u}(t)) = \frac{m}{2}\dot{u}^2(t) \quad \text{and} \quad V(u(t)) = \frac{k}{2}u^2(t),$$

and thus

$$I(u) = \int_{t_1}^{t_2} \big(\frac{m}{2}\dot{u}^2(t) - \frac{k}{2}u^2(t)\big)\, dt.$$

To motivate the expression $V(u(t)) = \frac{k}{2}u^2(t)$ for the potential energy, we use the definition of the potential energy as the total work required to move the mass from position $u = 0$ to position $u(t)$. The work to move the mass from position v to $v + \Delta v$ is equal to $kv\Delta v$ following the principle that work = force $\times$ displacement. The total work is thus

$$V(u(t)) = \int_0^{u(t)} kv\, dv = \frac{k}{2}u^2(t),$$

as announced.

To see how the equation $m\ddot{u} + ku = 0$ arises, we compute the derivative of $I(u+\epsilon w)$ with respect to ϵ and then set $\epsilon = 0$, where $w(x)$ is a perturbation satisfying $w(t_1) = w(t_2) = 0$. Direct computation based on moving $\frac{d}{d\epsilon}$ inside the integral, which is allowed since the limits of integration are fixed,

$$\begin{aligned}\frac{d}{d\epsilon}I(u + \epsilon w) &= \int_{t_1}^{t_2} \frac{d}{d\epsilon}\big(\frac{m}{2}\dot{u}\dot{u} + \epsilon m\dot{u}\dot{w} + \frac{m}{2}\epsilon^2\dot{w}\dot{w} - \frac{k}{2}u^2 - k\epsilon uw - \frac{k}{2}\epsilon^2 w^2\big)\, dt \\ &= \int_{t_1}^{t_2} \big(m\dot{u}\dot{w} - kuw\big)\, dt \quad \text{for } \epsilon = 0.\end{aligned}$$

Integrating by parts in the term $m\dot{u}\dot{w}$, we get

$$\int_{t_1}^{t_2} \big(m\ddot{u} + ku\big)w\, dt = 0,$$

for all $w(t)$ with $w(t_1) = w(t_2) = 0$. This implies that $m\ddot{u} + ku = 0$ in $[t_1, t_2]$, since $w(t)$ can vary arbitrarily in the interval (t_1, t_2), and we obtain the desired equation.

47.3 Pendulum with fixed support

We consider a pendulum in the form of a body of mass one attached to a weightless string of unit length fixed to the ceiling under the action of a vertical gravity force normalized to one. The action integral of the difference between kinetic and potential energy is given by

$$I(u) = \int_{t_1}^{t_2} \big(\frac{1}{2}\dot{u}^2(t) - (1 - \cos(u(t))\big)\,dt,$$

where $u(t)$ represents the angle of the pendulum in radians at time t, measured from the vertical position, see Fig. 47.2. The potential energy in this

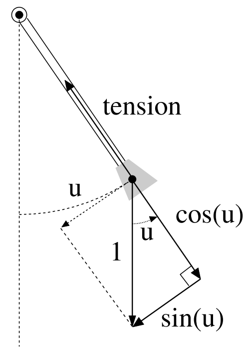


FIGURE 47.2. A pendulum.

case is equal to the work of lifting the mass from the bottom position to the level $(1-\cos(v))$, which is exactly equal to $(1-\cos(v))$ if the gravitational constant is normalized to one. Stationarity of the action integral requires that for all perturbations $w(t)$ satisfying $w(t_1) = w(t_2) = 0$, we have

$$0 = \frac{d}{d\epsilon}\int_{t_1}^{t_2} \big(\frac{1}{2}(\dot{u}+\epsilon\dot{w})^2(t) - (1 - \cos(u(t)+\epsilon w(t))\big)\,dt \quad \text{for } \epsilon = 0,$$

which gives as above

$$\int_{t_1}^{t_2} \big(\ddot{u} + \sin(u(t))\big) w\,dt = 0.$$

This yields the initial value problem

$$\begin{cases} \ddot{u} + \sin(u) = 0 & \text{for } t > 0 \\ u(0) = u_0,\ \dot{u}(0) = u_1, \end{cases} \tag{47.3}$$

where we added initial conditions for position and velocity.

The resulting differential equation $\ddot{u} = -\sin(u)$ is an expression of Newton's Law, since $\ddot{u}$ is the angular acceleration and $-\sin(u)$ is the restoring force in the angular direction. We conclude that the Principle of Least Action in the present case is a reformulation of Newton's Law.

If the angle of the pendulum stays small during the motion, then we can approximate $\sin(u)$ by u and obtain the linear equation $\ddot{u} + u = 0$, with solutions being linear combinations of $\sin(t)$ and $\cos(t)$.

47.4 Pendulum with moving support

We now generalize to a pendulum with a support that is subject to a prescribed motion. Consider thus a body of mass m swinging in a weightless string of length l that is attached to a support moving according to a given function $r(t) = (r_1(t), r_2(t))$ in a coordinate system with the x_1-axis horizontal and the x_2-axis vertical upward. Let $u(t)$ be the angle of the string at time t measured from the vertical.

The potential energy is again equal to the height of the body, measured from some reference position, times mg with g the gravitational constant. Thus, we may choose

$$V(u(t)) = mg(r_2(t) - l\cos(u)).$$

To express the kinetic energy, we need to take into account the motion of the support. The velocity of the body relative to the support is $(l\dot{u}\cos u, l\dot{u}\sin u)$, and the total velocity is thus $(\dot{r}_1(t) + l\dot{u}\cos u, \dot{r}_2(t) + l\dot{u}\sin u)$. The kinetic energy is $m/2$ times the square of the *modulus of the velocity*, and thus

$$T = \frac{m}{2}\left[(\dot{r}_1 + l\dot{u}\cos u)^2 + (\dot{r}_1 + l\dot{u}\sin u)^2\right].$$

Using the Principle of Least Action, we obtain the following equation:

$$\ddot{u} + \frac{g}{l}\sin u + \frac{\ddot{r}_1}{l}\cos u + \frac{\ddot{r}_2}{l}\sin u = 0, \tag{47.4}$$

together with initial conditions for $u(0)$ and $\dot{u}(0)$.

If the support is fixed with $\ddot{r}_1 = \ddot{r}_2 = 0$, then we recover the equation (47.3) setting $l = m = g = 1$.

47.5 The Principle of Least Action

We now consider a mechanical system that is described by a vector function $u(t) = (u_1(t), u_2(t))$. We may think of a system consisting of two bodies with positions given by the functions $u_1(t)$ and $u_2(t)$. The action integral is

$$I(u_1, u_2) = I(u) = \int_{t_1}^{t_2} L(u(t)\, dt,$$

where

$$L(u_1(t), u_2(t)) = L(u(t)) = T(\dot{u}(t)) - V(u(t))$$

is the difference of the kinetic energy $T(\dot{u}(t)) = T(\dot{u}_1(t), \dot{u}_2(t))$ and the potential energy $V(u(t)) = V(u_1(t), u_2(t))$. We refer to $L(u(t))$ as the *Lagrangean* of the state $u(t)$.

The Principle of Least Action states that the action integral is stationary at the true state $u(t)$ in the sense that for all perturbations $w_1(t)$ and $w_2(t)$ with $w_1(t_1) = w_2(t_1) = w_1(t_2) = w_2(t_2)$, we have for $\epsilon = 0$,

$$\begin{aligned}
&\frac{d}{d\epsilon} I(u_1 + \epsilon w_1, u_2) = 0\\
&\frac{d}{d\epsilon} I(u_1, u_2 + \epsilon w_2) = 0.
\end{aligned}$$

Assuming that

$$T(\dot{u}_1(t), \dot{u}_2(t)) = \frac{m_1}{2} \dot{u}_1^2(t) + \frac{m_2}{2} \dot{u}_2^2(t),$$

we obtain performing the differentiation with respect to ϵ as above and setting $\epsilon = 0$,

$$\begin{aligned}
&\int_{t_1}^{t_2} \big(m\dot{u}_1(t)\dot{w}_1(t) - \frac{\partial V}{\partial u_1}(u_1(t), u_2(t))w_1(t)\big)\, dt = 0,\\
&\int_{t_1}^{t_2} \big(m\dot{u}_2(t)\dot{w}_2(t) - \frac{\partial V}{\partial u_2}(u_1(t), u_2(t))w_2(t)\big)\, dt = 0.
\end{aligned}$$

Integrating by parts as above and letting w_1 and w_2 vary freely over (t_1, t_2), we obtain

$$\begin{aligned}
m\ddot{u}_1(t) &= -\frac{\partial V}{\partial u_1}(u_1(t), u_2(t)),\\
m\ddot{u}_2(t) &= -\frac{\partial V}{\partial u_2}(u_1(t), u_2(t)).
\end{aligned} \tag{47.5}$$

If we set

$$F_1 = -\frac{\partial V}{\partial u_1}, \qquad F_2 = -\frac{\partial V}{\partial u_2},$$

then we can write the equations derived from the Principle of Least Action as

$$\begin{aligned}
m\ddot{u}_1(t) &= F_1(u_1(t), u_2(t)),\\
m\ddot{u}_2(t) &= F_1(u_1(t), u_2(t)),
\end{aligned} \tag{47.6}$$

which can be viewed as Newton's Law if F_1 and F_2 are interpreted as forces.

47.6 Conservation of the total energy

Defining the *total energy*

$$E(u(t)) = T(\dot{u}(t)) + V(u(t))$$

as the sum of the kinetic and potential energies and using (47.5), we get

$$\begin{aligned}\frac{d}{dt}E(u(t)) &= m_1\dot{u}_1\ddot{u}_1 + m_2\dot{u}_2\ddot{u}_2 + \frac{\partial V}{\partial u_1}\dot{u}_1 + \frac{\partial V}{\partial u_2}\dot{u}_2\\ &= \dot{u}_1\big(m_1\ddot{u}_1 + \frac{\partial V}{\partial u_1}\big) + \dot{u}_2\big(m_2\ddot{u}_2 + \frac{\partial V}{\partial u_2}\big) = 0.\end{aligned}$$

We conclude that the total energy $E(u(t))$ is constant in time, that is the energy is *conserved*. Obviously, energy conservation is not a property of all systems, and thus the Principle of Least Action only applies to so called *conservative systems*, where the total energy is conserved. In particular, effects of *friction* are not present.

47.7 The double pendulum

We now consider a *double pendulum* consisting of two bodies of masses m_1 and m_2, where the first body of mass m_1 hangs on a weightless string of length l_1 attached to a fixed support and the second body of mass m_2 hangs on a weightless string of length l_2 attached to the first body. We shall now apply the Principle of Least Action to derive the equations of motion for this system.

To describe the state of the system, we use the angles $u_1(t)$ and $u_2(t)$ of the two bodies measured from vertical position.

We now seek expressions for the kinetic and potential energies of the system of the two bodies. The contributions from the second body is obtained from the expressions for a pendulum with moving support derived above if we set $(r_1(t), r_2(t)) = (l_1 \sin u_1, -l_1 \cos u_1)$.

The potential energy of the first pendulum is $-mgl_1 \cos u_1$ and the total potential energy is

$$V(u_1(t), u_2(t)) = -m_1 g l_1 \cos u_1(t) - m_2 g\left(l_1 \cos u_1(t) + l_2 \cos u_2(t)\right).$$

The total kinetic energy is obtained similarly adding the kinetic energies of the two bodies:

$$\begin{aligned}T(\dot{u}_1(t), \dot{u}_2(t)) = \frac{m_1}{2}l_1^2\dot{u}_1^2 + \frac{m_2}{2}\big[&(l_1\dot{u}_1\cos u_1 + l_2\dot{u}_2\cos u_2)^2\\ &+(l_1\dot{u}_1\sin u_1 + l_2\dot{u}_2\sin u_2)^2\big].\end{aligned}$$

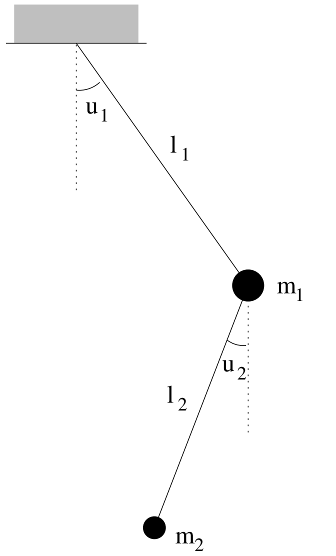


FIGURE 47.3. Double pendulum

Using the identities $\sin^2 u + \cos^2 u = 1$ and $\cos(u_1 - u_2) = \cos u_1 \cos u_2 + \sin u_1 \sin u_2$, we can rewrite this expression as

$$T = \frac{m_1}{2} l_1^2 \dot{u}_1^2 + \frac{m_2}{2}\left[l_1^2\dot{u}_1^2 + l_2^2\dot{u}_2^2 + 2l_1 l_2 \dot{u}_1 \dot{u}_2 \cos(u_1 - u_2)\right].$$

Applying the Principle of Least Action, we obtain the following system of equations for a double pendulum:

$$\begin{aligned} \ddot{u}_1 + \frac{m_2}{m_1+m_2}\frac{l_2}{l_1}[\ddot{u}_2\cos(u_2-u_1) - \dot{u}_2^2\sin(u_2-u_1)] + \frac{g}{l_1}\sin u_1 &= 0, \\ \ddot{u}_2 + \frac{l_1}{l_2}[\ddot{u}_1\cos(u_2-u_1) + \dot{u}_1^2\sin(u_2-u_1)] + \frac{g}{l_2}\sin u_2 &= 0. \end{aligned} \tag{47.7}$$

We note that if $m_2 = 0$, then the first equation is just the equation for a simple pendulum, and that if $\ddot{u}_1 = \dot{u}_1 = 0$, then the second equation is again the equation for a simple pendulum.

47.8 The two-body problem

We consider the *two-body* problem for a small mass orbiting around a heavy mass, such as the Earth moving around the Sun neglecting the influence of the other planets. We assume that the motion takes place in a plane and use polar coordinates (r, θ) with the origin at the center of the heavy mass to describe the position of the light body. Assuming that the heavy body is fixed, the action integral representing the difference between the kinetic and potential energy of the small mass is given by

$$\int_{t_1}^{t_2} \left(\frac{1}{2}\dot{r}^2 + \frac{1}{2}(\dot{\theta} r)^2 + \frac{1}{r}\right) dt \tag{47.8}$$

because the velocity is $(\dot{r}, r\dot{\theta})$ in the radial and angular directions respectively, and the gravity potential is $-r^{-1} = -\int_r^\infty s^{-2}\, ds$ corresponding to the work needed to move a particle of unit mass a distance r from the orbit center to infinity. The corresponding Euler-Lagrange equations are

$$\begin{cases} \ddot{r} - r\dot{\theta}^2 = -\frac{1}{r^2}, & t > 0, \\ \frac{d}{dt}(r^2\dot{\theta}) = 0, & t > 0, \end{cases} \tag{47.9}$$

which is a second order system to be complemented with initial values for position and velocity.

We construct the analytical solution of this system in a set of problems below, which may be viewed as a short course on Newton's *Principia Matematica*. We invite the reader to take this opportunity of getting on speaking terms with Newton himself.

47.9 Stability of the motion of a pendulum

The linearization of the equation for a pendulum at $\bar{u} \in \mathbb{R}$, $\ddot{u} + \sin(u) = 0$, is obtained by setting $u = \bar{u} + \varphi$ and noting that $\sin(u) \approx \sin(\bar{u}) + \cos(\bar{u})\varphi$. This leads to

$$0 = \ddot{u} + \sin(u) \approx \ddot{\varphi} + \sin(\bar{u}) + \cos(\bar{u})\varphi.$$

Assuming first that $\bar{u} = 0$, we obtain the following linearized equation for the perturbation φ,

$$\ddot{\varphi} + \varphi = 0, \tag{47.10}$$

with solution being a linear combination of $\sin(t)$ and $\cos(t)$. For example, if $\varphi(0) = \delta$ and $\dot{\varphi}(0) = 0$, then $\varphi(t) = \delta\cos(t)$, and we see that an initially small perturbation is kept small for all time: the pendulum stays close to the bottom position under small perturbations.

Setting next $\bar{u} = \pi$, we obtain

$$\ddot{\varphi} - \varphi = 0 \tag{47.11}$$

with the solution being a linear combination of $\exp(\pm t)$. Since $\exp(t)$ grows very quickly, the state $\bar{u} = \pi$ corresponding to the pendulum in the top position is *unstable*. A small perturbation will quickly develop into a large perturbation and the pendulum will move way from the top position.

We will return to the topic of this section in Chapter *Linearization and stability of initial value problems*

Chapter 47 Problems

47.1. Supply the missing details in the derivation of the equation for the pendulum. If the angle u stays small during the motion, then the simpler *linearized* model $\ddot{u} + u = 0$ may be used. Solve this equation analytically and compare with numerical results for the nonlinear pendulum equation to determine limits of validity of the linear model.

47.2. Carry out the details in the derivation of the equations for the pendulum with moving support and the double pendulum.

47.3. Study what happens for the double pendulum in the extreme cases, i.e. at zero and infinity, for the parameters m_1, m_2, l_1 and l_2.

47.4. Derive the second equation of motion for the double pendulum from the result for the pendulum with moving support by setting $(r_1(t), r_2(t)) = (l_1 \sin u_1, -l_1 \cos u_1)$.

47.5. Derive the equation of motion for a bead sliding on a frictionless plane vertical curve under the action of gravity.

47.6. In the foot-steps of Newton give an analysis and analytical solution of the two-body problem modeled by (47.9) through the following sequence of problems: (i) Prove that a stationary point of the action integral (47.8) satisfies (47.9). (ii) Prove that the total energy is constant in time. (iii) Introducing the change of variables $u = r^{-1}$, show that $\dot{\theta} = cu^2$ for c constant. Use this relation together with the fact that the chain rule implies that

$$\frac{dr}{dt} = \frac{dr}{du}\frac{du}{d\theta}\frac{d\theta}{dt} = -c\frac{du}{d\theta} \quad \text{and} \quad \ddot{r} = -c^2u^2\frac{d^2u}{d\theta^2}$$

to rewrite the system (47.9) as

$$\frac{d^2u}{d\theta^2} + u = c^{-2}. \tag{47.12}$$

Show that the general solution of (47.12) is

$$u = \frac{1}{r} = \gamma\cos(\theta - \alpha) + c^{-2},$$

where γ and α are constants. (iii) Prove that the solution is either an ellipse, parabola, or hyperbola. Hint: Use the fact that these curves can be described as the loci of points for which the ratio of the distance to a fixed point and to a fixed straight line, is constant. Polar coordinates are suitable for expressing this relation. (iv) Prove Kepler's three laws for planetary motion using the experience from the previous problem.

47.7. Study the linearizations of the double pendulum at $(u_1, u_2) = (0, 0)$ and $(u_1, u_2) = (\pi, \pi)$ and draw conclusions about stablity.

47.8. Attach an elastic string to a simple pendulum in some way and model the resulting system.

47.9. Compute solutions of the presented models numerically.

48
N-body systems*

> The reader will find no figures in this work. The methods which I set forth do not require either geometrical or mechanical reasonings, but only algebraic operations, subject to a regular and uniform rule of procedure. (Lagrange in *Méchanique Analytique*

48.1 Introduction

We shall now model systems of N bodies interacting through mechanical forces that result from springs and dashpots, see Fig. 48.2, or from gravitational or electrostatic forces. We shall use two different modes of description. In the first formulation, we describe the system through the coordinates of (the centers of gravity of) the bodies. In the second, we use the *displacements* of the bodies measured from an initial reference configuration. In the latter case, we also *linearize* under an assumption of small displacements to obtain a linear system of equations. In the first formulation, the initial configuration is only used to initialize the system and is "forgotten" at a later time in the sense that the description of the system only contains the present position of the masses. In the second formulation, the reference configuration is retrievable through the evolution since the unknown is the displacement from the reference position. The different formulations have different advantages and ranges of application.

48.2 Masses and springs

We consider the motion in $\mathbb{R}^3$ of a system of N bodies connected by a set of Hookean springs. For $i = 1, ..., N$, let the position at time t of body i be given by the vector function $u_i(t) = (u_{i1}(t), u_{i2}(t), u_{i3}(t))$, with $u_{ik}(t)$ denoting the x_k coordinate, $k = 1, 2, 3$, and suppose the mass of body i is m_i. Let body i be connected to body j with a Hookean spring of spring constant $k_{ij} \geq 0$ for $i, j = 1, ...N$. Some of the k_{ij} may be zero, which effectively means that there is no spring connection between body i and body j. In particular $k_{ii} = 0$. We assume to start with that the reference length of the spring corresponding to zero spring tension is equal to zero. This means that the spring forces are always attractive.

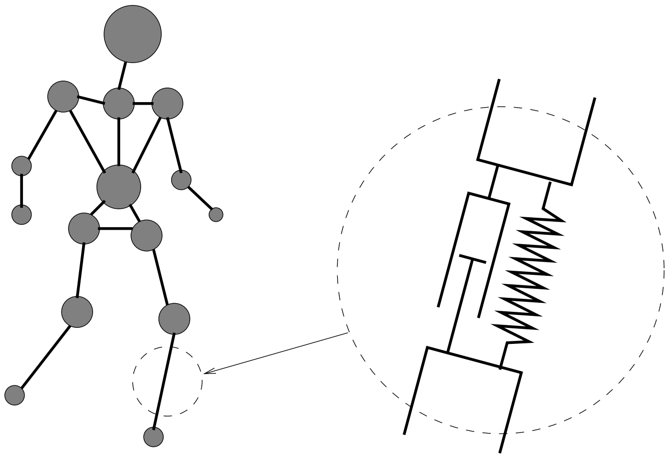


FIGURE 48.1. A typical system of masses, springs and dashpots in motion

We now derive the equations of motion for the mass-spring system using the Principle of Least Action. We assume first that the gravitational force is set to zero. The potential energy of the configuration $u(t)$ is given by

$$\begin{aligned} V(u(t)) &= \sum_{i,j=1}^{N} \frac{1}{2} k_{ij} |u_i - u_j|^2 \\ &= \sum_{i,j=1}^{N} \frac{1}{2} k_{ij} \big((u_{i1} - u_{j1})^2 + (u_{i2} - u_{j2})^2 + (u_{i3} - u_{j3})^2 \big), \end{aligned} \tag{48.1}$$

with the time dependence of the coordinates u_{ik} suppressed for readability. This is because the length of the spring connecting the body i and body j is equal to $|u_i - u_j|$, and the work to stretch the spring from zero length to length l is equal to $\frac{1}{2} k_{ij} l^2$.

The action integral is

$$I(u) = \int_{t_1}^{t_2} \big(\sum_{i=1}^{N} \frac{1}{2} m_i \big(\dot{u}_{i1}^2 + \dot{u}_{i2}^2 + \dot{u}_{i3}^2\big) - V(u(t))\big)\, dt,$$

and using the Principle of Least Action and the fact that

$$\frac{\partial V(u)}{\partial u_{ik}} = \sum_{j=1}^{N} k_{ij}(u_{ik} - u_{jk}),$$

we obtain the following system of equations of motion:

$$m_i \ddot{u}_{ik} = -\sum_{j=1}^{N} k_{ij}(u_{ik} - u_{jk}), \quad k = 1, 2, 3,\ i = 1, ..., N, \tag{48.2}$$

or in vector form

$$m_i \ddot{u}_i = -\sum_{j=1}^{N} k_{ij}(u_i - u_j),\ i = 1, ..., N, \tag{48.3}$$

together with initial conditions for $u_i(0)$ and $\dot{u}_i(0)$. We can view these equations as expressing Newton's Law

$$m_i \ddot{u}_i = F_i^s, \tag{48.4}$$

with the total spring force $F_i^s = (F_{i1}^s, F_{i2}^s, F_{i3}^s)$ acting on body i being equal to

$$F_i^s = -\sum_{j=1}^{N} k_{ij}(u_i - u_j). \tag{48.5}$$

Inclusion of gravity forces in the direction of the negative x_3 axis, adds a component $-m_i g$ to F_{i3}^s, where g is the gravitational constant.

The system (48.3) is linear in the unknowns $u_{ij}(t)$. If we assume that the reference length with zero spring force of the spring connecting body i and j is equal to $l_{ij} > 0$, then the potential changes to

$$V(u(t)) = \sum_{i,j=1}^{N} \frac{1}{2} k_{ij}(|u_i - u_j| - l_{ij})^2, \tag{48.6}$$

and the resulting equations of motion are no longer linear. Below, we shall consider a linearized form assuming $|u_i - u_j| - l_{ij}$ is small compared to l_{ij}.

48.3 The N-body problem

By tradition, a "N-body" problem refers to a system of N bodies in motion in $\mathbb{R}^3$ under the influence of mutual gravitational forces. An example is given by our solar system with 9 planets orbiting around the Sun, where we typically disregard moons, asteroids, and comets.

Let the position at time t of (the center of gravity of) body i be given by the vector function $u_i(t) = (u_{i1}(t), u_{i2}(t), u_{i3}(t))$, with $u_{ik}(t)$ denoting the x_k coordinate in $\mathbb{R}^3$, $k = 1, 2, 3$, and suppose the mass of body i is m_i. Newton's inverse square law of gravitation states that the gravitational force from the body j on the body i is given by

$$-\frac{\gamma m_i m_j}{|u_i(t) - u_j(t)|^2} \frac{u_i(t) - u_j(t)}{|u_i(t) - u_j(t)|} = -\gamma m_i m_j \frac{u_i(t) - u_j(t)}{|u_i(t) - u_j(t)|^3},$$

where γ is a gravitational constant. We thus obtain the following system of equations modeling the N-body problem:

$$m_i \ddot{u}_i = -\gamma m_i m_j \sum_{j \neq i} \frac{u_i - u_j}{|u_i(t) - u_j(t)|^3}, \tag{48.7}$$

together with initial conditions for $u_i(0)$ and $\dot{u}_i(0)$.

Alternatively, we may derive these equations using the Principle of Least Action using the gravity potential

$$V(u) = -\sum_{i,j=1,\, i \neq j}^{N} \frac{\gamma m_i m_j}{|u_i - u_j|},$$

and the fact that

$$\frac{\partial V}{\partial u_{ik}} = \sum_{j \neq i} \frac{\gamma m_i m_j}{|u_i - u_j|^3}(u_{ik} - u_{jk}). \tag{48.8}$$

The expression for the gravity potential is obtained by noticing that the work to bring body i from a distance r of body j to infinity is equal to

$$\int_r^\infty \frac{\gamma m_i m_j}{s^2}\, ds = \gamma m_i m_j [-\frac{1}{s}]_{s=r}^{s=\infty} = \frac{\gamma m_i m_j}{r}.$$

Notice the minus sign of the potential, arising from the fact that the body i loses potential energy as it approaches body j.

Analytical solutions are available only in the case of the 2-body problem. The numerical solution of for example the 10-body problem of our solar system is very computationally demanding in the case of long time simulation. As a result, the long time stability properties of our Solar system are unknown. For example, it does not seem to be known if eventually the Earth will change orbit with Mercury, Pluto will spin away to another galaxy, or some other dramatic event will take place.

The general result of existence guarantees a solution, but the presence of the stability factor $\exp(tL_f)$ brings the accuracy in long-time simulation seriously in doubt.

A nice picture of our solar system

48.4 Masses, springs and dashpots: small displacements

We now give a different description of the mass-spring system above. Let the initial position of body i, which is now chosen as reference position, be $a_i = (a_{i1}, a_{i2}, a_{i3})$, and let the actual position at time $t > 0$ be given by $a_i + u_i(t)$ where now $u_i(t) = (u_{i1}(t), u_{i2}(t), u_{i3}(t))$ is the *displacement* of body i from its reference position a_i.

The potential energy of the configuration $u(t)$ is given by

$$
\begin{aligned}
V(u(t)) &= \sum_{i,j=1}^{N} \frac{1}{2} k_{ij} \big(|a_i + u_i - (a_j + u_j)| - |a_i - a_j|\big)^2 \\
&= \frac{1}{2} k_{ij} \big(|a_i - a_j + (u_i - u_j)| - |a_i - a_j|\big)^2,
\end{aligned}
$$

assuming zero spring forces if the springs have the reference lengths $a_i - a_j$.

We now specialize to small displacements, assuming that $|u_i - u_j|$ is small compared to $|a_i - a_j|$. We then use that if $|b|$ is small compared to $|a|$, where $a, b \in \mathbb{R}^3$, then

$$
\begin{aligned}
|a+b| - |a| &= \frac{(|a+b| - |a|)(|a+b| + |a|)}{|a+b| + |a|} \\
&= \frac{|a+b|^2 - |a|^2}{|a+b| + |a|} = \frac{(a+b)\cdot(a+b) - a\cdot a}{|a+b| + |a|} \approx \frac{a \cdot b}{|a|}.
\end{aligned}
$$

Thus, if $|u_i - u_j|$ is small compared to $|a_i - a_j|$, then

$$
|a_i - a_j + (u_i - u_j)| - |a_i - a_j| \approx \frac{(a_i - a_j)\cdot(u_i - u_j)}{|a_i - a_j|},
$$

and we obtain the following approximation of the potential energy

$$
\hat{V}(u(t)) = \sum_{i,j=1}^{N} \frac{1}{2} k_{ij} \frac{\big((a_i - a_j)\cdot(u_i - u_j)\big)^2}{|a_i - a_j|^2}.
$$

Using the Principle of Least Action we thus obtain the following linearized system of equations

$$m_i \ddot{u}_{ik} = -\sum_{j=1}^{N} \frac{k_{ij}(a_i - a_j)\cdot(u_i - u_j)}{|a_i - a_j|^2}(a_{ik} - a_{jk}), \quad k = 1,2,3,\, i = 1,...,N,$$

or in vector form

$$m_i \ddot{u}_i = -\sum_{j=1}^{N} \frac{k_{ij}(a_i - a_j)\cdot(u_i - u_j)}{|a_i - a_j|^2}(a_i - a_j), \quad i = 1,...,N. \qquad (48.9)$$

together with initial conditions for $u_i(0)$ and $\dot{u}_i(0)$. We can view these equations as expressing Newton's Law

$$m_i \ddot{u}_i = F_i^s, \quad i + 1,...,N, \qquad (48.10)$$

with the spring force F_i^s acting on body i given by

$$F_i^s = -\sum_{j=1}^{N} b_{ij} e_{ij},$$

where

$$e_{ij} = \frac{a_i - a_j}{|a_i - a_j|}$$

is the normalized vector connecting a_j and a_i, and

$$b_{ij} = k_{ij} e_{ij} \cdot (u_i - u_j). \qquad (48.11)$$

48.5 Adding dashpots

A *dashpot* is a kind of shock absorber which may be thought of as consisting of a piston that moves inside a cylinder filled with oil or some other viscous fluid, see Fig. 48.2. As the piston moves, the flow of the fluid past the piston creates a force opposite to the motion, which we assume is proportional to the velocity with the constant of proportionality representing the coefficient of *viscosity* of the dashpot.

We now expand the above mass-spring model to include springs and dashpots coupled in parallel. For each pair of nodes i and j, we let k_{ij} and μ_{ij} be the coefficients of a spring and a dashpot coupled in parallel, with $k_{ij} = 0$ and $\mu_{ij} = 0$ if the spring or dashpot is absent, and in particular $k_{ii} = \mu_{ii} = 0$. The dashpot force F_i^d acting on body i will then be given by

$$F_i^d = -\sum_{j=1}^{N} d_{ij} e_{ij},$$

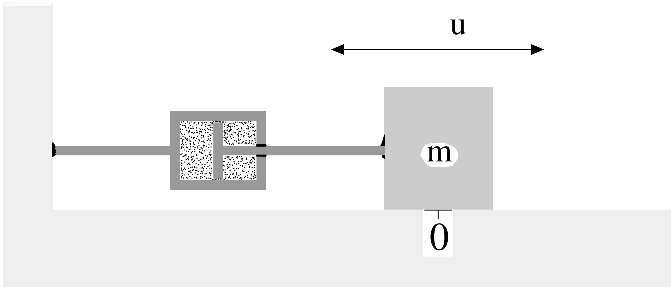


FIGURE 48.2. Cross section of a dashpot connected to a mass.

where

$$d_{ij} = \mu_{ij} e_{ij} \cdot (\dot{u}_i - \dot{u}_j). \tag{48.12}$$

To get this result, we use the fact that

$$e_{ij} \cdot (\dot{u}_i - \dot{u}_j) e_{ij}$$

is the projection of $\dot{u}_i - \dot{u}_j$ onto e_{ij}. We thus assume that the dashpot reacts with a force that is proportional to the projection of $\dot{u}_i - \dot{u}_j$ onto the direction $a_i - a_j$.

This leads to the linearized mass-spring-dashpot model:

$$m_i \ddot{u}_i = F_i^s + F_i^d, \quad i = 1, ..., N, \tag{48.13}$$

together with initial conditions for $u_i(0)$ and $\dot{u}_i(0)$. We can write these equations as a system in the form

$$M\ddot{u} + D\dot{u} + Ku = 0, \tag{48.14}$$

with constant coefficient matrices matrices M, D and K, where u is a $3N$-vector listing all the components u_{ik}. The matrix M is diagonal with the masses m_i as entries, and D and K are symmetric positive semi-definite (see the problem section).

A system with dashpots is not conservative, since the dashpots consume energy, and therefor cannot be modeled using the Principle of Least Action.

The linear system (48.14) models a wide range of phenomena and can be solved numerically with appropriate solvers. We return to this issue below. We now consider the simplest example of one mass connected to the origin with a spring and a dashpot in parallel.

48.6 The linear oscillator

We now consider the simplest example consisting of one body of mass 1 connected to one end of a Hookean spring connected to the origin with the motion taking place along the x_1-axis. Assuming the spring has zero length at zero tension, the system is described by

$$\begin{cases} \ddot{u} + ku = 0 & \text{for } t > 0, \\ u(0) = u_0, \ \dot{u}(0) = \dot{u}_0. \end{cases} \tag{48.15}$$

with $u(t)$ denoting the x_1 coordinated of the body at time t, and u_0 and $\dot{u}_0$ given initial conditions. The solution is given by

$$u(t) = a\cos(\sqrt{k}t) + b\sin(\sqrt{k}t) = \alpha\cos(\sqrt{k}(t-\beta)), \tag{48.16}$$

where the constants a and b, or α and β, are determined by the initial conditions. We conclude that the motion of the mass is periodic with *frequency* $\sqrt{k}$ and *phase shift* β and *amplitude* α, depending on the initial data. This model is referred to as the *linear oscillator*. The solution is periodic with period $\frac{2\pi}{\sqrt{k}}$, and the *time scale* is similar.

48.7 The damped linear oscillator

Adding a dashpot in parallel with the spring in the model above gives the model of a damped linear oscillator

$$\begin{cases} \ddot{u} + \mu\dot{u} + ku = 0, & \text{for } t > 0, \\ u(0) = u_0, \quad \dot{u}(0) = \dot{u}_0. \end{cases} \tag{48.17}$$

In the case $k = 0$, we obtain the model

$$\begin{cases} \ddot{u} + \mu\dot{u} = 0 & \text{for } t > 0, \\ u(0) = u_0, \quad \dot{u}(0) = \dot{u}_0, \end{cases} \tag{48.18}$$

with the solution

$$u(t) = -\frac{\dot{u}_0}{\mu}\exp(-\mu t) + u_0 + \frac{\dot{u}_0}{\mu}.$$

We see that the mass approaches the fixed position $u = u_0 + \frac{\dot{u}_0}{\mu}$ determined by the initial data as t increases to infinity. The time scale is of size $\frac{1}{\mu}$.

The characteristic polynomial equation for the full model $\ddot{u}+\mu\dot{u}+ku = 0$, is

$$r^2 + \mu r + kr = 0.$$

Completing the square we can write the characteristic equation in the form

$$(r + \frac{\mu}{2})^2 = \frac{\mu^2}{4} - k = \frac{1}{4}(\mu^2 - 4k). \tag{48.19}$$

If $\mu^2 - 4k > 0$, then there are two real roots $-\frac{1}{2}(\mu \pm \sqrt{\mu^2 - 4k})$, and the solution $u(t)$ has the form (see the Chapter The exponential function),

$$u(t) = ae^{-\frac{1}{2}(\mu+\sqrt{\mu^2-4k})t} + be^{-\frac{1}{2}(\mu-\sqrt{\mu^2-4k})t},$$

with the constants a and b determined by the initial conditions. In this case, the viscous damping of the dashpot dominates over the spring force, and the solution converges exponentially to a rest position, which is equal to $u = 0$ if $k > 0$. The fastest time scale is again of size $\frac{1}{\mu}$.

If $\mu^2 - 4k < 0$, then we introduce the new variable $v(t) = e^{\frac{\mu t}{2}}u(t)$, with the objective of transforming the characteristic equation (48.19) into an equation of the form $s^2 + (k - \frac{\mu^2}{4}) = 0$. Since $u(t) = e^{-\frac{\mu t}{2}}v(t)$, we have

$$\dot{u}(t) = \frac{d}{dt}\big(e^{-\frac{\mu t}{2}}v(t)\big) = (\dot{v} - \frac{\mu}{2}v)e^{-\frac{\mu t}{2}},$$
$$\ddot{u}(t) = (\ddot{v} - \mu\dot{v} + \frac{\mu^2}{4})e^{-\frac{\mu t}{2}},$$

and thus the differential equation $\ddot{u} + \mu\dot{u} + ku = 0$ is transformed into

$$\ddot{v} + (k - \frac{\mu^2}{4})v = 0,$$

with the solution $v(t)$ being a linear combination of $\cos(\frac{t}{2}\sqrt{4k - \mu^2})$ and $\sin(\frac{t}{2}\sqrt{4k - \mu^2})$. Transforming back to the variable $u(t)$ we get the solution formula

$$u(t) = ae^{-\frac{1}{2}\mu t}\cos(\frac{t}{2}\sqrt{4k - \mu^2}) + be^{-\frac{1}{2}\mu t}\sin(\frac{t}{2}\sqrt{4k - \mu^2}).$$

The solution again converges to the zero rest position as time passes if $\mu > 0$, but now it does so in an oscillatory fashion. Now two time scales appear: a time scale of size $\frac{1}{\mu}$ for the exponential decay and a time scale $1/\sqrt{k - \mu^2/4}$ of the oscillations.

Finally, in the limit case $\mu^2 - 4k = 0$ the solution $v(t)$ of the corresponding equation $\ddot{v} = 0$ is given by $v(t) = a + bt$, and thus

$$u(t) = (a + bt)e^{-\frac{1}{2}\mu t}.$$

This solution exhibits initial linear growth and eventually converges to a zero rest position as time tends to infinity. We illustrate the three possible behaviors in Fig. 48.3.

48.8 Extensions

We have above studied systems of bodies interacting through Hookean springs, linear dashpots and gravitational forces. We can generalize to systems of non-linear springs, dashpots, and other mechanical devices like

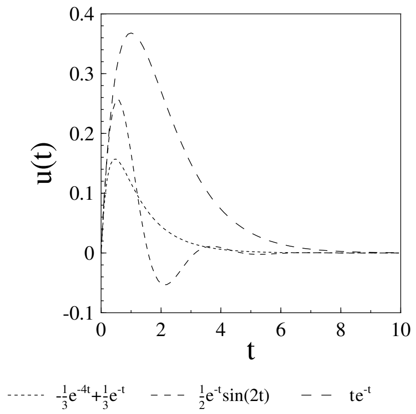


FIGURE 48.3. Three solutions of the mass-spring-dashpot model (48.17) satisfying the initial conditions $u(0) = 0$ and $\dot{u}(0) = 1$. The first solution corresponds to $\mu = 5$ and $k = 4$, the second to $\mu = 2$ and $k = 5$, and the third to $\mu = 2$ and $k = 1$.

springs reacting to changes of angles between the bodies, or other forces like electrostatic forces. In this way, we can model very complex systems from macroscopic scales of galaxies to microscopic molecular scales. For example, electrostatic forces are related to potentials of the form

$$V^e(u) = \pm c \sum_{i,j=1}^{N} \frac{q_i q_j}{|u_i - u_j|}$$

where q_i is the charge of body i and c is a constant, and thus have a form similar to that of gravitational forces.

In particular, models for molecular dynamics take the form of N-body systems interacting through electrostatic forces and forces modeled by various springs reacting to bond lengths and bond angles between the atoms. In these applications, N may be of the order 10^4 and the smallest time scale of the dynamics may be of size 10^{-14} related to very stiff bond length springs. Needless to say, simulations with such models may be very computationally demanding and is often out of reach with present day computers. For more precise information, we refer to the survey article *Molecular modeling of proteins and mathematical prediction of protein structure*, SIAM REV. (39), No 3, 407-460, 1997, by A. Neumair.

Chapter 48 Problems

48.1. Verify the solution formulas for the three solutions shown in Fig. 48.3.

48.2. Write down the model (48.2) in a simple case of a system with a few bodies.

48.3. Derive the equations of motion with the potential (48.6).

48.4. Generalize the mass-spring-dashpot model to arbitrary displacements.

48.5. Generalize the mass-spring model to different non-linear springs.

48.6. Model the vertical motion of a floating buoy. Hint: use that by Archimedes' Principle, the upward force on a cylindrical vertical bouy from the water is proportional to the immersed depth of the bouy.

48.7. Prove that the matrices D and K in (48.14) are symmetric positive semi-definite.

49
The Crash model*

On October 24, 1929, people began selling their stocks as fast as they could. Sell orders flooded market exchanges. On a normal day, only 750-800 members of the New York Stock Exchange started the Exchange. However, there were 1100 members on the floor for the morning opening. Furthermore, the Exchange directed all employees to be on the floor since there were numerous margin calls and sell orders placed overnight and extra telephone staff was arranged at the members' boxes around the floor. The Dow Jones Industrial Index closed at 299 that day. October 29 was the beginning of the Crash. Within the first few hours the stock market was open, prices fell so far as to wipe out all the gains that had been made in the previous year. The Dow Jones Industrial Index closed at 230. Since the stock market was viewed as the chief indicator of the American economy, public confidence was shattered. Between October 29 and November 13 (when stock prices hit their lowest point) over $30 billion disappeared from the American economy. It took nearly twenty-five years for many stocks to recover. (www.arts.unimelb.edu.au/amu/ucr/student/1997/Yee/1929.htm)

49.1 Introduction

Why did the Wall fall on November 9 1989? Why did the Soviet Union dissolve in January 1992? Why did the Stock market collapse in October 1929 and 1987? Why did Peter and Mary break up last Fall after 35 years of marriage? What caused the September 11 attack? Why does the flow in the river go from orderly laminar to chaotic turbulent at a certain specific point? All the situations behind these questions share a common feature: Nothing particularly dramatic preceded the sudden transition from stable to unstable, and in each case the rapid and dramatic change away from normality came as big surprise to almost everyone.

We now describe a simple mathematical model that shows the same behavior: the solution stays almost constant for a long time and then quite suddenly the solution explodes.

We consider the following initial value problem for a system of two ordinary differential equations: find $u(t) = (u_1(t), u_2(t))$ such that

$$\begin{cases} \dot{u}_1 + \epsilon u_1 - \lambda u_2 u_1 = \epsilon & t > 0, \\ \dot{u}_2 + 2\epsilon u_2 - \epsilon u_1 u_2 = 0 & t > 0, \\ u_1(0) = 1, u_2(0) = \kappa\epsilon, \end{cases} \tag{49.1}$$

where ϵ is a small positive constant of size say 10^{-2} or smaller and λ and κ are positive parameters of moderate size ≈ 1. If $\kappa = 0$, then the solution $u(t) = (1, 0)$ is constant in time, which we view as the *base solution.* In general, for $\kappa > 0$, we think of $u_1(t)$ as a primary part of solution with initial value $u_1(0) = 1$, and $u_2(t)$ as a small secondary part with an initial value $u_2(0) = \kappa\epsilon$ that is small because ϵ is small. Both components $u_1(t)$ and $u_2(t)$ will correspond to physical quantities that are non-negative and $u_1(0) = 1$ and $u_2(0) = \kappa\epsilon \geq 0$.

49.2 The simplified growth model

The system (49.1) models an interaction between a primary quantity $u_1(t)$ and a secondary quantity $u_2(t)$ through the terms $-\lambda u_1 u_2$ and $-\epsilon u_2 u_1$. If we keep just these terms, we get a simplified system of the form

$$\begin{cases} \dot{w}_1(t) = \lambda w_1(t) w_2(t) & t > 0, \\ \dot{w}_2(t) = \epsilon w_2(t) w_1(t) & t > 0, \\ w_1(0) = 1, \quad w_2(0) = \kappa\epsilon. \end{cases} \tag{49.2}$$

We see that the coupling terms are *growth terms* in the sense that both the equation $\dot{w}_1(t) = \lambda w_1(t) w_2(t)$ and $\dot{w}_2(t) = \epsilon w_2(t) w_1(t)$ say that $\dot{w}_1(t)$ and $\dot{w}_2(t)$ are positive if $w_1(t) w_2(t) > 0$. In fact, the system (49.1) always blow up for $\kappa > 0$ because the two components propel each other to infinity as t increases in the sense that the right hand sides get bigger with $w_1(t) w_2(t)$ and this increases the growth rates $\dot{w}_1(t)$ and $\dot{w}_2(t)$, which in turn makes $w_1 w_2(t)$ even bigger, and so on towards blow up, see Fig. 49.1.

We can study the blow up in (49.2) analytically assuming for simplicity that $\lambda = \kappa = 1$. In this case, it turns out that the two components $w_1(t)$ and

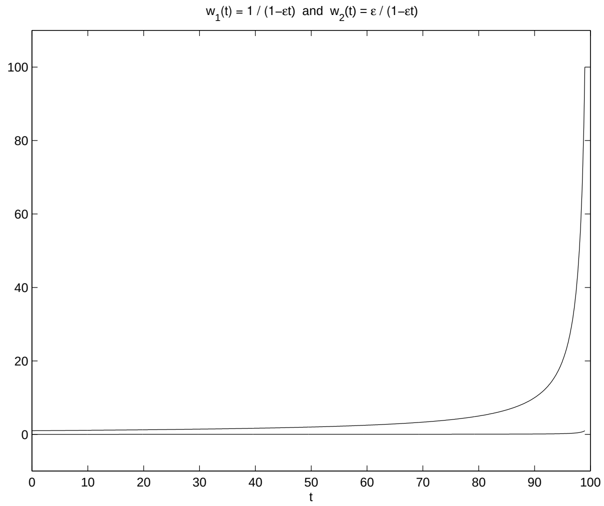


FIGURE 49.1. Solution of simpified growth model

$w_2(t)$ for all t are coupled by the relation $w_2(t) = \epsilon w_1(t)$, that is $w_2(t)$ is always the same multiple of $w_1(t)$. We check this statement by first verifying that $w_2(0) = \epsilon w_1(0)$ and then by dividing the two equations to see that $\dot{w}_2(t)/\dot{w}_1(t) = \epsilon$. So, $\dot{w}_2(t) = \epsilon \dot{w}_1(t)$, that is $w_2(t) - w_2(0) = \epsilon w_1(t) - \epsilon w_2(0)$, and we get the desired conclusion $w_2(t) = \epsilon w_1(t)$ for $t > 0$. Inserting this relation into the first equation of (49.2), we get

$$\dot{w}_1(t) = \epsilon w_1^2(t) \qquad \text{for } t > 0,$$

which can be written as

$$-\frac{d}{dt}\frac{1}{w_1(t)} = \epsilon \qquad \text{for } t > 0.$$

Recalling the initial condition $w_1(0) = 1$, we get

$$-\frac{1}{w_1(t)} = \epsilon t - 1 \qquad \text{for } t \geq 0,$$

which gives the following solution formula in the case $\lambda = \kappa = 1$:

$$w_1(t) = \frac{1}{1 - \epsilon t}, \quad w_2(t) = \frac{\epsilon}{1 - \epsilon t} \qquad \text{for } t \geq 0. \tag{49.3}$$

This formula shows that the solution tends to infinity as t increases towards $1/\epsilon$, that is, the solution explodes at $t = 1/\epsilon$. We notice that the time of blow up is $1/\epsilon$, and that the *time scale* before the solution starts to increase noticeably, is of size $\frac{1}{2\epsilon}$, which is a long time since ϵ is small. Thus, the solution changes very slowly for a long time and then eventually blows up quite a bit more rapidly, see Fig. 49.1.

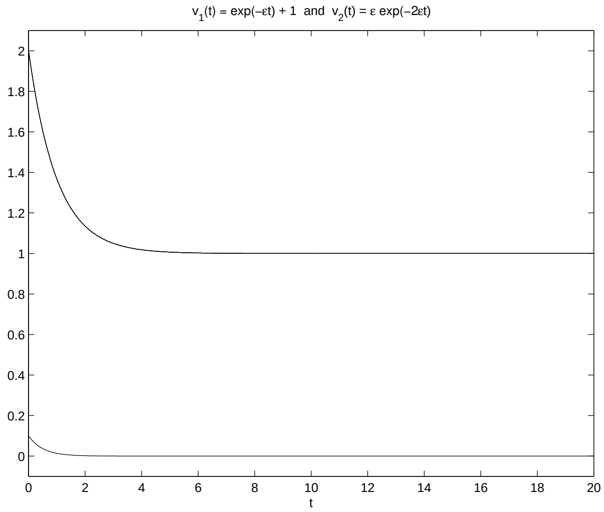


FIGURE 49.2. Solution of simpified decay model

49.3 The simplified decay model

On the other hand, if we forget about the growth terms, we get another simplified system:

$$\begin{cases} \dot{v}_1 + \epsilon v_1 = \epsilon \qquad t > 0, & t > 0, \\ \dot{v}_2 + 2\epsilon v_2 = 0 & t > 0, \\ v_1(0) = 1 + \delta, \quad v_2(0) = \kappa\epsilon, \end{cases} \tag{49.4}$$

where we have also introduced a small perturbation δ in $v_1(0)$. Here the two terms ϵv_1 and $2\epsilon v_2$ are so called *dissipative* terms that cause the solution $v(t)$ to return to the base solution $(1, 0)$ regardless of the perturbation, see Fig. 49.2. This is clear in the equation $\dot{v}_2 + 2\epsilon v_2 = 0$ with solution $v_2(t) = v_2(0)\exp(-2\epsilon t)$, which decays to zero as t increases. Rewriting the equation $\dot{v}_1 + \epsilon v_1 = \epsilon$ as $\dot{V}_1 + \epsilon V_1 = 0$, setting $V_1 = v_1 - 1 = \exp(-\epsilon t)$, we find that $v_1(t) = \delta\exp(-\epsilon t) + 1$, and thus $v_1(t)$ approaches 1 as t increases. We summarize: the solution $(v_1(t), v_2(t))$ of (49.6) satisfies

$$v_1(t) = \delta\exp(-\epsilon t) + 1 \to 1, \quad v_2(t) = \kappa\epsilon\exp(-2\epsilon t) \to 0, \quad \text{as } t \to \infty.$$

We say that (49.6) is a *stable* system because the solution always returns from $(1+\delta, \kappa\epsilon)$ to the base solution $(1, 0)$ independently of the perturbation $(\delta, \kappa\epsilon)$ of $(v_1(0), v_2(0))$.

We note that the time scale is again of size $1/\epsilon$, because of the presence of the factors $\exp(-\epsilon t)$ and $\exp(-2\epsilon t)$.

49.4 The full model

We can now sum up: The real system (49.1) is a combination of the unstable system (49.2) that includes only the growth terms only and whose the solution always blows up, and the stable system (49.6) that excludes the growth terms. We shall see that depending on the size of $\lambda\kappa$ the unstable or stable feature will take over. In Fig. 49.3 and Fig. 49.4, we show different solutions for different values of the parameters λ and κ with different initial values $u(0) = (u_1(0), u_2(0)) = (1, \kappa\epsilon)$. We see that if $\lambda\kappa$ is sufficiently large, then the solution $u(t)$ eventually blows up after a time of size $1/\epsilon$, while if $\lambda\kappa$ is sufficiently small, then the solution $u(t)$ returns to the base solution $(1,0)$ as t tends to infinity.

Thus, there seems to be a *threshold value* for $\lambda\kappa$ above which the initially disturbed solution eventually blows up and below which the initially disturbed solution returns to the base solution. We can view κ as a measure of the size if the initial disturbance, because $u_2(0) = \kappa\epsilon$. Further, we can view the factor λ as a quantitative measure of the *coupling* between the growth components $u_2(t)$ and $u_1(t)$ through the growth term $\lambda u_1 u_2$ in the evolution equation for u_1.

Our main conclusion is that if the initial disturbance times the coupling is sufficiently large, then the system will blow up. Blow up thus requires both the initial disturbance and the coupling to be sufficiently large. A large initial disturbance will not cause blow up unless there some coupling. A strong coupling will not cause blow up unless there is an initial disturbance.

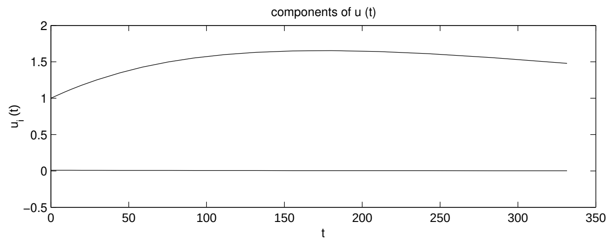


FIGURE 49.3. Return to the base solution if $\lambda\kappa$ is small enough

We now investigate the qualitative behavior of (49.1) in a little more detail. We see that $\dot{u_1}(0)/u_1(0) = \lambda\kappa\epsilon$, while $\dot{u_2}(0)/u_2(0) = -\epsilon$, which shows that initially $u_1(t)$ grows and $u_2(t)$ decays at relative rates of size ϵ. Now, $u_1(t)$ will continue to grow as long as $\lambda u_2(t) > \epsilon$, and further $u_2(t)$ will start to grow as soon as $u_1(t) > 2$. Thus, if $u_1(t)$ manages to become larger than 2, before $u_2(t)$ has decayed below ϵ/λ, then both components will propel each other to cause a blow up to infinity. This happens if $\lambda\kappa$ is above a certain threshold.

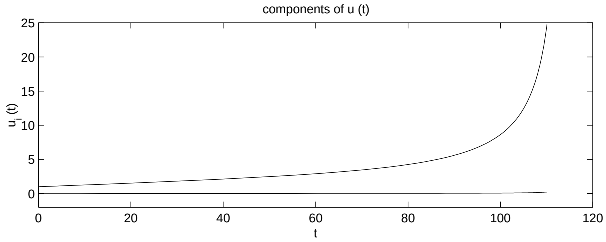


FIGURE 49.4. Blow up if $\lambda\kappa$ is large enough

We notice that the time scale for significant changes in both u_1 and u_2 is of size ϵ^{-1}, because the growth rates are of size ϵ. This conforms with the experience from the simplified models. The scenario is thus that the primary part $u_1(t)$ grows slowly starting from 1 at a rate of size ϵ and the secondary part $u_2(t)$ decays slowly at a rate of size ϵ^2, over a time of size $1/\epsilon$. If $\lambda\kappa$ is above a certain threshold, then $u_1(t)$ reaches the value 2, at which point $u_2(t)$ starts to grow and eventually blow up follows on a somewhat shorter time scale. If $u_1(t)$ does not reach the value 2 in time, then $(u_1(t), u_2(t))$ returns to the base solution $(1, 0)$ as t increases.

We hope the presented scenario is quite easy to grasp intuitively, and conforms with every-day experiences of quit sudden blow-up, as a result of an accumulation of small events over a long period.

We can give the Crash model very many interpretations in real life, such as

- stock market (u_1 stock prize of big company, u_2 stock prize of small innovative company),
- chemical reaction (u_1 main reactant, u_2 catalyst),
- marriage crisis (u_1 main discontent, u_2 small irritation factor),
- spread of infection (u_1 infected people, u_2 amount of germs),
- symbiosis (u_1 main organism, u_2 small parasite),
- population model (u_1 rabbits, u_2 vitalizing carrots),

and many others.

In particular, the model describes an essential aspect of the process of transition from laminar to turbulent flow in for example a pipe. In this case u_1 represents a flow component in the direction of the pipe and u_2 represents a small perturbation of the flow in the transversal direction. The time to explosion corresponds to the time it takes for the flow to go turbulent starting as laminar flow at the inlet. In the famous experiment of Reynolds from 1888, ink is injected at the inlet of a transparent pipe and the observer can follow the streamline traced by the ink, which forms a straight line in the laminar part and then successively becomes more and more wavy until it breaks down to completely turbulent flow at some distance from the inlet. The distance to breakdown varies with the flow speed and viscosity and perturbations resulting from e.g. roughness of the surface of the pipe or a heavy-weight truck passing by at some distance from the experimental set-up.

Chapter 49 Problems

49.1. Develop the indicated applications of the Crash model.

49.2. Solve the full system (49.1) numerically for various values of λ and κ and try to pin down the threshold value of $\lambda\kappa$.

49.3. Develop a *Theory of Capitalism* based on (49.1) as a simple model of the economy in a society, with u_1 representing the value of a basic resource like land, and u_2 some venture capital related to the exploitation of new technology, with $(1, 0)$ a base solution without the new technology, and with the coeffcient λ of the u_1u_2 term in the first equation representing the positive interplay between base and new technology, and the terms ϵu_i representing stabilizing effects of taxes for example. Show that the possible pay-off $u_1(t) - u_1(0)$ of a small investment $u_2(0) = \kappa\epsilon$ may be large, and that an exploding economy may result if $\lambda\kappa$ is large enough. Show that no growth is possible if $\lambda = 0$. Draw some conclusions from the model coupled to for example the role of the interest rate for controlling the economy.

49.4. Interprete (49.1) as a simple model of a stock market with two stocks, and discuss scenarios of overheating. Extend to a model for the world stock market, and predict the next crash.

49.5. Consider the linear model

$$\begin{aligned} \dot\varphi_1 + \epsilon\varphi_1 - \lambda\varphi_2 &= 0 \qquad t > 0, \\ \dot\varphi_2 + \epsilon\varphi_2 &= 0 \qquad t > 0, \\ \varphi_1(0) = 0, \quad \varphi_2(0) &= \kappa\epsilon, \end{aligned} \tag{49.5}$$

which is obtained from (49.1) by setting $\varphi_1 = u_1 - 1$ and $\varphi_2 = u_2$ and replacing $u_1\varphi_2$ by φ_2 assuming u_1 is close to 1. Show that the solution of (49.5) is given by

$$\varphi_2(t) = \kappa\epsilon\exp(-\epsilon t), \quad \varphi_1(t) = \lambda\kappa\epsilon t\exp(-\epsilon t).$$

Conclude that

$$\frac{\varphi_1(\frac{1}{\epsilon})}{\varphi_2(0)} = \lambda\frac{\exp(-1)}{\epsilon},$$

and make an interpretation of this result.

49.6. Expand the Crash model (49.1) to

$$\begin{aligned}
\dot{u}_1 + \epsilon u_1 - \lambda u_1 u_2 + \mu_1 u_2^2 &= \epsilon \qquad t > 0,\\
\dot{u}_2 + 2\epsilon u_2 - \epsilon u_2 u_1 + \mu_2 u_1^2 &= 0 \qquad t > 0,\\
u_1(0) = 1, \quad u_2(0) &= \kappa\epsilon,
\end{aligned} \tag{49.6}$$

with decay terms $\mu_1 u_2^2$ and $\mu_2 u_1^2$, where μ_1 and μ_2 are positive coefficients. (a) Study the stabilizing effect of such terms numerically. (b) Seek to find values of μ_1 and μ_2, so that the corresponding solution starting close to $(1,0)$ shows an intermittent behaviour with repeated periods of blow up followed by a decay back to a neigborhood of $(1,0)$. (c) Try to find values of μ_1 and μ_2 so that multiplication of the first equation with a positive multiple of u_1 and the second by u_2, leads to bounds on $|\epsilon u_1(t)|^2$ and $|u_2(t)|^2$ in terms of initial data. Hint: Try for example $\mu_1 \approx 1/\epsilon$, and $\mu_2 \approx \epsilon^2$.

49.7. Study the initial value problem $\dot{u} = f(u)$ for $t > 0$, $u(0) = 0$, where $f(u) = \lambda u - u^3$, with different values of $\lambda \in \mathbb{R}$. Relate the time-behaviour of $u(t)$ to the set of solutions $\bar{u}$ of $f(u) = 0$, that is, $\bar{u} = 0$ if $\lambda \leq 0$, and $\bar{u} = 0$ or $\bar{u} = \pm\sqrt{\lambda}$ if $\lambda > 0$. Study the linearized models $\dot{\varphi} - \lambda\varphi + 3\bar{u}^2\varphi = 0$ for the different $\bar{u}$. Study the behaviour of the solution assuming $\lambda(t) = t - 1$.

49.8. Study the model

$$\begin{aligned}
\dot{w}_1 + w_1 w_2 + \epsilon w_1 &= 0, \qquad t > 0,\\
\dot{w}_2 - \epsilon w_1^2 + \epsilon w_2 &= -\gamma\epsilon, \qquad t > 0,
\end{aligned} \tag{49.7}$$

with given initial data $w(0)$, where γ is a parameter and $\epsilon > 0$. This problem admits the stationary "trivial branch" solution $\bar{w} = (0, -\gamma)$ for all γ. If $\gamma > \epsilon$, then also $\bar{w} = (\pm\sqrt{\gamma - \epsilon}, -\epsilon)$ is a stationary solution. Study the evolution of the solution for different values of γ. Study the corresponding linearized problem, linearized at $\bar{w}$.

50 Electrical circuits*

We can scarcely avoid the conclusion that light consists in the transverse undulations of the same medium which is the cause of electric and magnetic phenomena. (Maxwell 1862)

50.1 Introduction

There is an analogy between models of masses, dashpots, and springs in mechanics and models of *electrical circuits* involving *inductors*, *resistors*, and *capacitors* respectively. The basic model for an electrical circuit, see Fig. 50.1, with these three components has the form

$$L\ddot{q}(t) + R\dot{q}(t) + \frac{q(t)}{C} = f(t), \quad \text{for} \quad t > 0, \tag{50.1}$$

together with initial conditions for q and $\dot{q}$. Here $f(t)$ represents an applied *voltage* and $q(t)$ is a primitive function of the *current* $i(t)$. The equation (50.1) says that the applied voltage $f(t)$ is equal to the sum of the *voltage drops* $L\frac{di}{dt}$, Ri and u/C across the inductor, resistor, and capacitor respectively, where L, R and C are the coefficients of *inductance*, *resistance* and *capacitance*. Note that the integral $q(t)$ of the current represents the *charge*.

The system (50.1), which is referred to as an LCR-circuit, takes the same form as the mass-dashpot-spring system (48.17), and the discussion above concerning the case $f(t) = 0$ applies to the LCR-circuit. In the absence of a resistor, a non-zero solution oscillates between extreme states with the charge $|u(t)|$ in the capacitor maximal and $\dot{u} = 0$ over an intermediate state

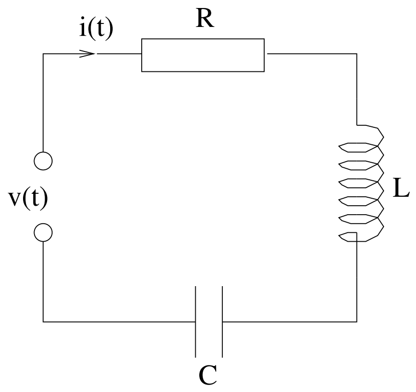


FIGURE 50.1. A circuit with inductor, resistor and capacitor.

with zero capacitor charge and $|\dot{u}|$ maximal. This is analogous to the mass-spring system with the potential energy corresponding to the capacitor and the velocity corresponding to $\dot{u}$. Further, the influence of a resistor is analogous to that of a dash pot, causing a damping of the oscillations.

We now proceed to describe the components of the electrical circuit in some more detail and show how complex circuits may be constructed combining the components in series or parallel in different configurations.

50.2 The components: inductors, resistors and capacitors

The voltage drop $v(t)$ across a capacitor satisfies

$$v(t) = \frac{q(t)}{C},$$

where $q(t)$ is the *charge* defined by

$$q(t) = \int_0^t i(t)\,dt,$$

assuming $q(0) = 0$, where $i(t)$ is the current and the constant C is the capacitance. Differentiating, we have

$$i(t) = \dot{q}(t) = C\dot{v}(t).$$

The voltage drop across a resistor is according to Ohm's Law $v(t) = Ri(t)$, where the constant R is the resistance. Finally, the voltage drop across an inductor is

$$v(t) = L\frac{di}{dt}(t),$$

where the constant L is the inductance.

50.3 Building circuits. Kirchhoff's laws

By joining the components of inductors, resistors and capacitors by electrical wires, we can build *electrical circuits*, with the wires joined at *nodes*. A closed loop in the circuit is a series of wires connecting components and nodes that leads back to the node of departure. To model the circuit, we use the two Kirchhoff laws:

- the sum of all currents entering a node is zero (first law),
- the sum of the voltage drops around any closed loop of the circuit is zero (second law).

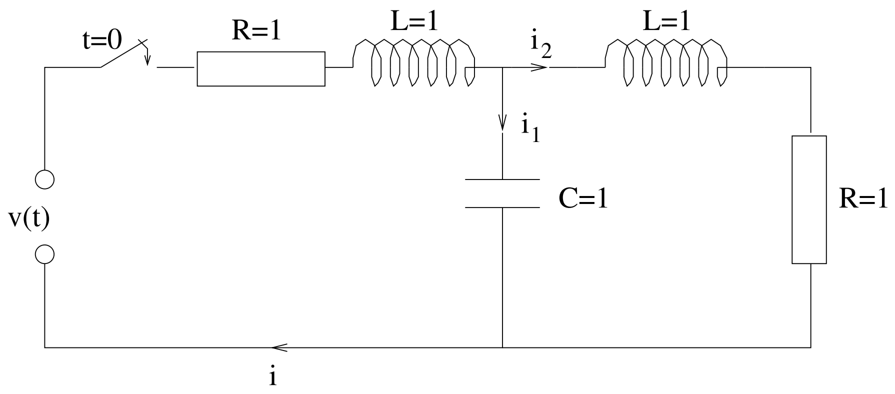


FIGURE 50.2. A circuit with two loops

EXAMPLE 50.1. We consider the following circuit consisting of two loops, see Fig. 50.3, and assume that $v(t) = 10$. Suppose that the switch is turned on at $t = 0$, with the charge in the capacitor being zero at $t = 0$ and $i_1(0) = i_2(0) = 0$. By Kirchhoffs second law applied to the two closed loops, we have

$$i + \frac{di}{dt} + \int_0^t i_1(s)\,ds = 10,$$
$$i_2 + \frac{i_2}{dt} - \int_0^t i_1(s)\,ds = 0.$$

Inserting Kirchhoffs first law stating that $i = i_1 + i_2$ into the first equation, and eliminating $i_2 + \frac{i_2}{dt}$ using the second equation, we get

$$i_1 + \frac{di_1}{dt} + 2\int_0^t i_1(s)\,ds = 10, \tag{50.2}$$

which upon differentiation gives the following second order equation,

$$\frac{d^2 i_1}{dt^2} + \frac{di_1}{dt} + 2i_1 = 0,$$

with characteristic equation $r^2 + r + 2 = 0$. Completing the square, we get the equation $(r + \frac{1}{2})^2 + \frac{7}{4} = 0$, and thus using the initial condition $i_1(0) = 0$, we have

$$i_1(t) = c\exp(-\frac{t}{2})\sin(\frac{t\sqrt{7}}{2}),$$

with c a constant. Inserting into (50.2) gives $c = \frac{20}{\sqrt{7}}$, and we have determined the current $i_1(t)$ as a function of time. We can now solve for i_2 using the second loop equation.

50.4 Mutual induction

Two inductors in different loops of the circuit may be coupled through *mutual inductance*, for instance by letting the inductors share an iron core. We give an example in Fig. 50.4.

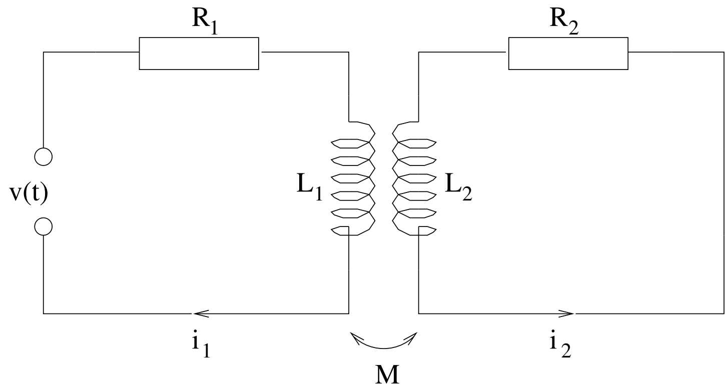


FIGURE 50.3. A circuit with mutual inductance

Kirchhoffs second law applied to each of the circuits now takes the form

$$R_1 i_1 + L_1 \frac{di_1}{dt} + M\frac{di_2}{dt} = v(t),$$
$$R_2 i_2 + L_1 \frac{di_2}{dt} + M\frac{di_1}{dt} = 0,$$

where L_1 and L_2 are the inductances in the two circuits and M is the coefficient of mutual inductance.

Chapter 50 Problems

50.1. Design circuits with the indicated components. Study resonance phenomena, and amplifiers.

50.2. Study the charging of a capacitor through a resistor.

50.3. Derive the effective resistance of n resistors coupled (a) in parallel, (b) in series. Do the same for inductors and capacitors.

51
String theory*

It's because physicists dream of a unified theory: a single mathematical framework in which all fundamental forces and units of matter can be described together in a manner that is internally consistent and consistent with current and future observation. And it turns out that having extra dimensions of space makes it possible to build candidates for such a theory. Superstring theory is a possible unified theory of all fundamental forces, but superstring theory requires a 10 dimensional spacetime, or else bad quantum states called ghosts with unphysical negative probabilities become part of the spectrum (http://superstringtheory.com/experm/exper5.html).

51.1 Introduction

We now study a couple of basic mass-spring models with different response characteristics depending on the geometry. These simple models lead to separable equations in the phase plane. The models can generalized and coupled into systems of high complexity.

Consider a horizontal elastic string attached at $(-1, 0$ and $(1, 0)$ in a $x-y$-plane with the y-axis vertical downward, and with a body of mass 1 attached to the string at its midpoint, see Fig. 51.1. To describe the dynamics of this system, we seek the vertical force required to displace the midpoint of the string the vertical distance y. Pythagoras theorem implies that the length of half the string after displacement is equal to $\sqrt{1+y^2}$, see Fig. 51.1. The elongation is thus $\sqrt{1+y^2}-1$ and assuming that the tension

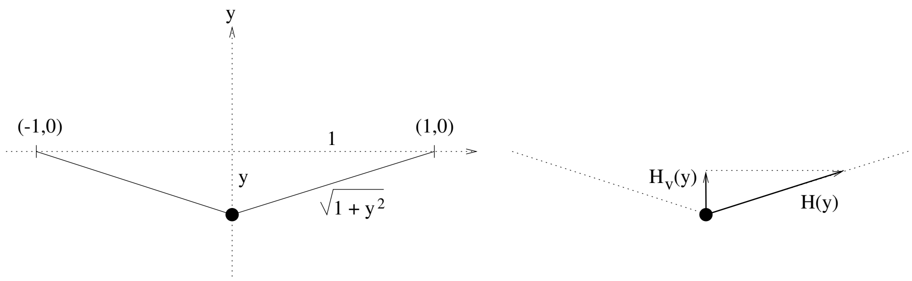


FIGURE 51.1. A horizontal elastic string **fix fig: y-axis downward and x-axis**

in the string before displacement is H and that the string is linear elastic with constant 1, we find that the tension of the string after displacement is equal to $H + (\sqrt{1+y^2} - 1)$. By similarity, the vertical component $H_v(y)$ of the string force is

$$H_v(y) = (H + \sqrt{1+y^2} - 1)\frac{y}{\sqrt{1+y^2}},$$

and the total downward force $f(y)$ required to pull the spring downward the distance y is thus

$$f(y) = 2H_v(y), \tag{51.1}$$

where the factor 2 comes from the fact that the upward force has contributions from both sides of the string, see Fig. 51.1. Using Newton's Law, we thus obtain the following model for the mass-spring system

$$\ddot{y} + f(y) = 0,$$

if we neglect gravity forces.

51.2 A linear system

We now assume that y so small that we can replace $\sqrt{1+y^2}$ by 1, and assuming that $H = \frac{1}{2}$, we then obtain the linear harmonic oscillator model:

$$\ddot{y} + y = 0, \tag{51.2}$$

with the solution $y(t)$ being a linear combination of $\sin(t)$ and $\cos(t)$. For example, if $y(0) = \delta$ with δ small and $\dot{y}(0) = 0$, then the solution is $y(t) = \delta\cos(t)$ corresponding to small vibrations around the rest position $y = 0$.

51.3 A soft system

We next assume that $H = 0$ so that the string is without tension with $y = 0$. By Taylor's theorem for y small (with an error proportional to y^4),

$$\sqrt{1+y^2} \approx 1 + \frac{y^2}{2},$$

and we obtain assuming y to be small

$$f(y) \approx \frac{y^3}{\sqrt{1+y^2}} \approx y^3,$$

and are thus led to the model

$$\ddot{y} + y^3 = 0.$$

In this model the restoring force y^3 is much smaller than the y in the linear model for y small, and thus corresponds to a "soft" system.

51.4 A stiff system

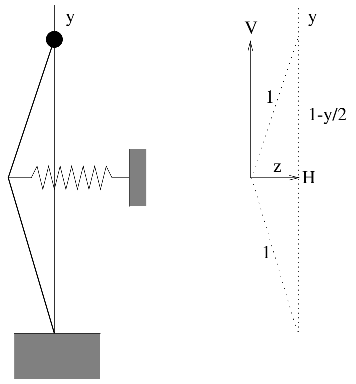


FIGURE 51.2. A "stiff" bar-spring system

We consider now a system of two nearly vertical bars each of length one connected by a frictionless joint with the lower bar fixed to the ground with a frictionless joint and with a body of mass 1 at the top of the upper bar, see Fig. 51.4. Let y be the vertical displacement downward of the weight from the top position with the bars fully vertical, and let z be the corresponding elongation of the horizontal spring. The corresponding horizontal spring force is $H = z$ assuming the spring is Hookean with zero natural length and spring constant equal to 1. With V denoting the vertical component of the bar force, momentum balance gives, see Fig. 51.4,

$$Vz = H(1 - \frac{y}{2}) = z(1 - \frac{y}{2}),$$

and thus the vertical force reaction on the body from the spring-bar system is given by

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

We see that the vertical force from the spring-bar system in this system is almost constant and equal to 1 for small y. Thus the system reacts with an almost constant response, which contrasts to the linear response y and the cubic response y^3 met above. In the present context, we may refer to this response as being "stiff" in contrast to the more or less soft response of the above systems. We are thus led to the following model in the present case

$$\ddot{y} + (1 - \frac{y}{2}) = 0, \tag{51.3}$$

where we assume that $0 \leq y \leq 1$. In particular, we may consider the initial conditions $y(0) = 0$, and $\dot{y}(0) = y_0 > 0$ and follow the evolution until $y(t)$ reaches the value 1 or 0.

51.5 Phase plane analysis

We rewrite the second order equation $\ddot{y} + f(y) = 0$ as the first order system

$$\dot{v} + f(u) = 0, \quad \dot{u} = v$$

by setting $u = y$ and $v = \dot{y}$. The coresponding phase plane equation reads

$$\frac{du}{dv} = -\frac{v}{f(u)},$$

which is a separable equation

$$f(u)\,du = -v\,dv$$

with solution curves in the $u - v$-plane satisfying

$$F(u) + \frac{v^2}{2} = C,$$

where $F(u)$ is a primitive function of $f(u)$. With $f(u) = u$ as in the linear case, the phase plane curves are circles,

$$u^2 + v^2 = 2C.$$

In the soft case with $f(u) = u^3$, the phase plane curves are given by

$$\frac{u^4}{4} + \frac{v^2}{2} = C,$$

which represent a sort of ellipses. In the stiff case with $f(u) = (1 - \frac{u}{2})$, the phase-plane curves are given by

$$-(1 - \frac{u}{2})^2 + \frac{v^2}{2} = C,$$

which represent hyperbolas.

Chapter 51 Problems

51.1. Compare the three systems from the point of view of efficiency as catapults, assuming the system can be "loaded" by applying a force of a certain maximal value and distance of action. Hint: Compute the work to load the system.

51.2. Design other mass-spring systems, for instance by coupling the elementary systems considered in series and parallel, and find their corresponding mathematical models. Solve the systems numerically.

51.3. Add gravity force to the above systems.

51.4. Develop an analog of the stiff system above of the form $\ddot{y} + (1 - y)$ if $0 \le y \le 1$ and $\ddot{y} - (1 + y) = 0$ if $-1 \le y \le 0$ allowing y to take arbitrary values in $-1 \le y \le 1$. Hint: mirror the given system.

51.5. Visualize the indicated phase-plane plots.

52
Piecewise linear approximation

> The results, however, of the labour and invention of this century are not to be found in a network of railways, in superb bridges, in enormous guns, or in instantaneous communication. We must compare the social state of the inhabitants of the country with what it was. The change is apparent enough. The population is double what it was a century back; the people are better fed and better housed, and comforts and even luxuries that were only within the reach of the wealthy can now be obtained by all classes alike.......But with these advantages there are some drawbacks. These have in many cases assumed national importance, and it has become the province of the engineer to provide a remedy. (Reynolds, 1868)

52.1 Introduction

Approximating a complicated function to arbitrary accuracy by "simpler" functions is a basic tool of applied mathematics. We have seen that piecewise polynomials are very useful for this purpose, and that is why approximation by piecewise polynomials plays a very important role in several areas of applied mathematics. For example, the *Finite Element Method* FEM is an extensively used tool for solving differential equations that is based on piecewise polynomial approximation, see the Chapters FEM for two-point boundary value problems and FEM for Poisson's equation.

In this chapter, we consider the problem of approximating a given real-valued function $f(x)$ on an interval $[a, b]$ by piecewise linear polynomials on a subdivision of $[a, b]$. We derive basic error estimates for interpolation with piecewise linear polynomials and we consider an application to least squares approximation.

52.2 Interpolation on $[0, 1]$

Let $f : [0, 1] \to \mathbb{R}$ be a given Lipschitz continuous function. Consider the function $\pi f : [0, 1] \to \mathbb{R}$ defined by

$$\pi f(x) = f(0)(1 - x) + f(1)x = f(0) + (f(1) - f(0))x.$$

Clearly, $\pi f(x)$ is a *linear* function in x,

$$\pi f(x) = c_0 + c_1 x,$$

where $c_0 = f(0)$, $c_1 = f(1) - f(0)$, and $\pi f(x)$ *interpolates* $f(x)$ at the end-points 0 and 1 of the interval $[0, 1]$, by which we mean that πf takes the same values as f at the end-points, i.e.

$$\pi f(0) = f(0), \quad \pi f(1) = f(1).$$

We refer to $\pi f(x)$ as a linear *interpolant* of $f(x)$ that interpolates $f(x)$ at the end-points of the interval $[0, 1]$.

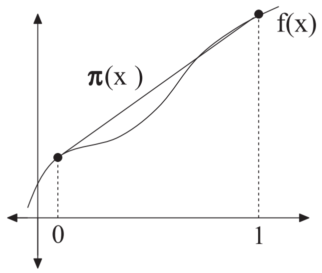


FIGURE 52.1. The linear interpolant πf of a function f.

We now study the *interpolation error* $f(x) - \pi f(x)$ for $x \in [0, 1]$. Before doing so we get some perspective on the space of linear functions on $[0, 1]$ to which the interpolant πf belongs.

52.3 The vector space of linear functions on $[0, 1]$

We let $\mathcal{P} = \mathcal{P}(0, 1)$ denote the set of first order (linear) polynomials

$$p(x) = c_0 + c_1 x,$$

defined for $x \in [0,1]$, where the real numbers c_0 and c_1 are the coefficients of p. We recall that two polynomials $p(x)$ and $q(x)$ in $\mathcal{P}$ may be added to give a new polynomial $p+q$ in $\mathcal{P}$ defined by $(p+q)(x) = p(x) + q(x)$, and that a polynomial $p(x)$ in $\mathcal{P}$ may be multiplied by a scalar α to give a polynomial αp in $\mathcal{P}$ defined by $(\alpha p)(x) = \alpha p(x)$. Adding two polynomials is carried out by adding their coefficients, and multiplying a polynomial by a real number is carried out by multiplying the coefficients by the real number.

We conclude that $\mathcal{P}$ is a vector space where each vector is a particular first order polynomial $p(x) = c_0 + c_1 x$ determined by the two real numbers c_0 and c_1. As a basis for $\mathcal{P}$ we may choose $\{1, x\}$. To see this, we note that each $p \in \mathcal{P}$ can be uniquely expressed as a linear combination of 1 and x: $p(x) = c_0 + c_1 x$, and we may refer to the pair (c_0, c_1) as the coordinates of the polynomial $p(x) = c_0 + c_1 x$ with respect to the basis $\{1, x\}$. For example, the coordinates of the polynomial $p(x) = x$ with respect to the basis $\{1, x\}$, are $(0, 1)$, right? Since there are two basis functions, we say that the dimension of the vector space $\mathcal{P}$ is equal to two.

We now consider an alternative basis $\{\lambda_0, \lambda_1\}$ for $\mathcal{P}$ consisting of the two functions λ_0 and λ_1 defined

$$\lambda_0(x) = 1 - x, \quad \lambda_1(x) = x.$$

Each of these functions takes the value 0 at one end-point and the value 1 at the other end-point, namely

$$\lambda_0(0) = 1,\ \lambda_0(1) = 0, \qquad \text{and } \lambda_1(0) = 0, \lambda_1(1) = 1.$$

See Fig. 52.2.

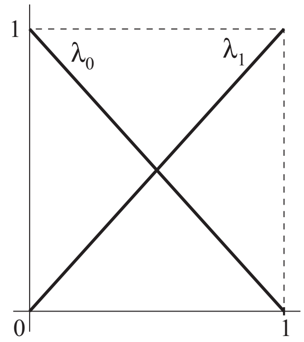


FIGURE 52.2. The basis functions λ_0 and λ_1.

Any polynomial $p(x) = c_0 + c_1 x$ in $\mathcal{P}$ can be expressed as a linear combination of the functions $\lambda_0(x)$ and $\lambda_1(x)$, i.e.

$$\begin{aligned} p(x) = c_0 + c_1 x = c_0(1-x) + (c_1 + c_0)x = c_0\lambda_0(x) + (c_1 + c_0)\lambda_1(x) \\ = p(0)\lambda_0(x) + p(1)\lambda_1(x), \end{aligned}$$

A very nice feature of these functions is that the coefficients $p(0)$ and $p(1)$ are the values of $p(x)$ at $x = 0$ and $x = 1$. Moreover, λ_0 and λ_1 are linearly independent, since if

$$a_0\lambda_0(x) + a_1\lambda_1(x) = 0 \quad \text{for } x \in [0,1],$$

then setting $x = 0$ and $x = 1$ shows that $a_1 = a_0 = 0$. We conclude that $\{\lambda_0, \lambda_1\}$ is a basis for $\mathcal{P}$.

In particular, we can express the interpolant $\pi f \in \mathcal{P}$ in the basis $\{\lambda_0, \lambda_1\}$ as follows:

$$\pi f(x) = f(0)\lambda_0(x) + f(1)\lambda_1(x), \tag{52.1}$$

where the end-point values $f(0)$ and $f(1)$ appear as coefficients.

52.3.1 The interpolation error

We want to estimate the interpolation error $f(x) - \pi f(x)$ for $x \in [0,1]$. We prove that

$$|f(x) - \pi f(x)| \leq \frac{1}{2}x(1-x) \max_{y\in[0,1]} |f''(y)|, \quad x \in [0,1]. \tag{52.2}$$

Since (convince yourself!)

$$0 \leq x(1-x) \leq \frac{1}{4} \quad \text{for } x \in [0,1],$$

we can state the interpolation error estimate in the form

$$\max_{x\in[0,1]} |f(x) - \pi f(x)| \leq \frac{1}{8} \max_{y\in[0,1]} |f''(y)|. \tag{52.3}$$

This estimate states that the maximal value of the interpolation error $|f(x)-\pi f(x)|$ over $[0,1]$ is bounded by a constant times the maximum value of the second derivative $|f''(y)|$ over $[0,1]$, i.e. to the degree of concavity or convexity of f, or the amount that f curves away from being linear, see Fig. 52.3.

To prove (52.2), we fix x in $(0,1)$ and use Taylor's theorem to express the values $f(0)$ and $f(1)$ in terms of $f(x)$, $f'(x)$, $f''(y_0)$ and $f''(y_1)$ where $y_0 \in (0,x)$ and $y_1 \in (x,1)$. This gives

$$\begin{aligned} f(0) &= f(x) + f'(x)(-x) + \frac{1}{2}f''(y_0)(-x)^2, \\ f(1) &= f(x) + f'(x)(1-x) + \frac{1}{2}f''(y_1)(1-x)^2. \end{aligned} \tag{52.4}$$

Substituting the Taylor expansions (52.4) into (52.1) and using the identities

$$\begin{aligned} \lambda_0(x) + \lambda_1(x) &= (1-x) + x \equiv 1, \\ (-x)\lambda_0(x) + (1-x)\lambda_1(x) &= (-x)(1-x) + (1-x)x \equiv 0, \end{aligned} \tag{52.5}$$

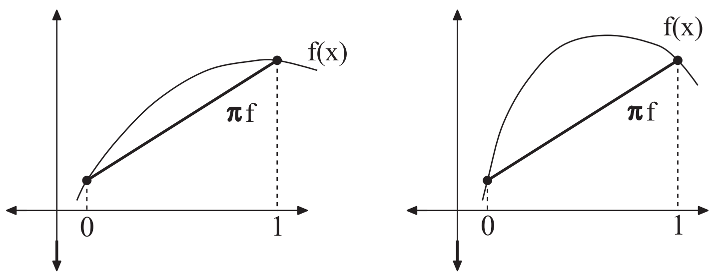


FIGURE 52.3. The error of a linear interpolant depends on the size of $|f''|$, which measures the degree that f curves away from being linear. Notice that the error of the linear interpolant of the function on the right is much larger than of the linear interpolant of the function on the left and the function on the right has a larger second derivative in magnitude.

we obtain the *error representation*

$$f(x) - \pi f(x) = -\frac{1}{2}\big(f''(y_0)(-x)^2(1-x) + f''(y_1)(1-x)^2 x\big).$$

Using the identity $(-x)^2(1-x)+(1-x)^2 x = x(1-x)(x+1-x) = x(1-x)$ gives (52.2),

$$|f(x) - \pi f(x)| \leq \frac{1}{2}x(1-x) \max_{y\in[0,1]} |f''(y)| \leq \frac{1}{8} \max_{y\in[0,1]} |f''(y)|. \qquad (52.6)$$

Next, we prove the following estimate for the error in the first derivative,

$$|f'(x) - (\pi f)'(x)| \leq \frac{x^2 + (1-x)^2}{2} \max_{y\in[0,1]} |f''(y)|, \quad x \in [0, 1]. \qquad (52.7)$$

Since $0 \leq x^2 + (1-x)^2 \leq 1$ for $x \in [0, 1]$,

$$\max_{x\in[0,1]} |f'(x) - (\pi f)'(x)| \leq \frac{1}{2} \max_{y\in[0,1]} |f''(y)|.$$

We illustrate in Fig. 52.4.

To prove (52.7), we differentiate (52.1) with respect to x (note that the x-dependence is carried by $\lambda_0(x)$ and $\lambda_1(x)$) and use (52.4) together with the obvious identities

$$\lambda_0'(x) + \lambda_1'(x) = -1 + 1 \equiv 0,$$
$$(-x)\lambda_0'(x) + (1-x)\lambda_1'(x) = (-x)(-1) + (1-x) \equiv 1.$$

This gives the error representation:

$$f'(x) - (\pi f)'(x) = -\frac{1}{2}\big(f''(y_0)(-x)^2(-1) + f''(y_1)(1-x)^2\big),$$

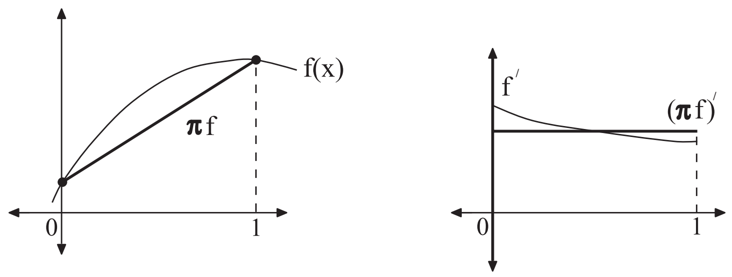


FIGURE 52.4. The derivative of a linear interpolant of f approximates the derivative of f. We show f and the linear interpolant πf on the left and their derivatives on the right.

where again $y_0 \in (0,x)$ and $y_1 \in (x,1)$. This proves the desired result.

Finally, we prove an estimate for $|f(x) - \pi f(x)|$ using only the first derivative f'. This is useful when the second derivative f'' does not exist. The Mean Value theorem implies

$$f(0) = f(x) + f'(y_0)(-x), \quad f(1) = f(x) + f'(y_1)(1-x), \tag{52.8}$$

where $y_0 \in [0,x]$ and $y_1 \in [x,1]$. Substituting into (52.1), we get

$$|f(x) - \pi f(x)| = |f'(y_0)x(1-x) - f'(y_1)(1-x)x| \leq 2x(1-x) \max_{y\in[0,1]} |f'(y)|.$$

Since $2x(1-x) \leq \frac{1}{2}$ for $0 \leq x \leq 1$, we thus find that

$$\max_{x\in[0,1]} |f(x) - \pi f(x)| \leq \frac{1}{2} \max_{y\in[0,1]} |f'(y)|.$$

We summarize in the following theorem.

Theorem 52.1 *The linear polynomial $\pi f \in \mathcal{P}(0,1)$, which interpolates the given function $f(x)$ at $x = 0$ and $x = 1$, satisfies the following error bounds:*

$$\begin{aligned}
\max_{x\in[0,1]} |f(x) - \pi f(x)| &\leq \frac{1}{8} \max_{y\in[0,1]} |f''(y)|, \\
\max_{x\in[0,1]} |f(x) - \pi f(x)| &\leq \frac{1}{2} \max_{y\in[0,1]} |f'(y)|, \\
\max_{x\in[0,1]} |f'(x) - (\pi f)'(x)| &\leq \frac{1}{2} \max_{y\in[0,1]} |f''(y)|.
\end{aligned} \tag{52.9}$$

The corresponding estimates for an arbitrary interval $I = [a,b]$ of length $h = b - a$ takes the following form, where of course $\mathcal{P}(a,b)$ denotes the set of linear functions on $[a,b]$. Observe how the length $h = b-a$ of the interval enters, with the factor h^2 in the estimate for $f(x) - \pi f(x)$ with f'', and h in the estimate for $f'(x) - (\pi f)'(x)$.

Theorem 52.2 *The linear polynomial $\pi f \in \mathcal{P}(a,b)$, which interpolates the given function $f(x)$ at $x = a$ and $x = b$, satisfies the following error bounds:*

$$\begin{aligned} \max_{x\in[a,b]} |f(x) - \pi f(x)| &\leq \frac{1}{8} \max_{y\in[a,b]} |h^2 f''(y)|, \\ \max_{x\in[a,b]} |f(x) - \pi f(x)| &\leq \frac{1}{2} \max_{y\in[a,b]} |h f'(y)|, \\ \max_{x\in[a,b]} |f'(x) - (\pi f)'(x)| &\leq \frac{1}{2} \max_{y\in[a,b]} |h f''(y)|, \end{aligned} \tag{52.10}$$

where $h = b - a$.

If we define the *maximum norm* over $I = [a,b]$ by

$$\|v\|_{L_\infty(I)} = \max_{x\in[a,b]} |v(x)|,$$

then we can state (52.9) as follows

$$\begin{aligned} \|f - \pi f\|_{L_\infty(I)} &\leq \frac{1}{8} \|h^2 f''\|_{L_\infty(I)}, \\ \|f - \pi f\|_{L_\infty(I)} &\leq \frac{1}{2} \|h f'\|_{L_\infty(I)}, \\ \|f' - (\pi f)'\|_{L_\infty(I)} &\leq \frac{1}{2} \|h f''\|_{L_\infty(I)}. \end{aligned} \tag{52.11}$$

Below we shall use an analog of this estimate with the $L_\infty(I)$-norm replaced by the $L_2(I)$-norm.

52.4 The vector space of piecewise linear continuous functions on an interval

For a given interval $I = [a,b]$, we let $a = x_0 < x_1 < x_2 < \cdots < x_N = b$ be a *partition* of I into N sub-intervals $I_i = (x_{i-1}, x_i)$ of length $h_i = x_i - x_{i-1}$, $i = 1, ..., N$. We denote by $h(x)$ the *mesh function* defined by $h(x) = h_i$ for $x \in I_i$ and we use $\mathcal{T}_h = \{I_i\}_{i=1}^N$ to denote the set of intervals or *mesh* or partition.

We introduce the vector space V_h of continuous piecewise linear functions on the mesh $\mathcal{T}_h$. A function $v \in V_h$ is linear on each subinterval I_i and is continuous on $[a,b]$. Adding two functions in V_h or multiplying a function in V_h by a real number gives a new function in V_h, and thus V_h is indeed a vector space. We show an example of such a function in Fig. 53.2.

We now present a particularly important basis for V_h that consists of the *hat functions* or *nodal basis functions* $\{\varphi_i\}_{i=0}^N$ illustrated in Fig. 52.5. The hat-function $\varphi_i(x)$ is a function in V_h satisfying

$$\varphi_i(x_j) = 1 \quad \text{if } j = i, \quad \varphi_i(x_j) = 0 \quad \text{if } j \neq i.$$

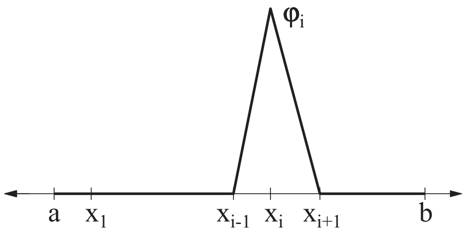


FIGURE 52.5. The hat function φ_i associated to node x_i.

and is given by the formula:

$$\varphi_i(x) = \begin{cases} 0, & x \notin [x_{i-1}, x_{i+1}], \\ \dfrac{x - x_{i-1}}{x_i - x_{i-1}}, & x \in [x_{i-1}, x_i], \\ \dfrac{x - x_{i+1}}{x_i - x_{i+1}}, & x \in [x_i, x_{i+1}]. \end{cases}$$

The basis functions φ_0 and φ_N associated to the boundary nodes x_0 and x_N look like "half hats". Observe that each hat function $\varphi_i(x)$ is defined on the whole interval $[a, b]$ and takes the value zero outside the interval $[x_{i-1}, x_{i+1}]$ (or $[a, x_1]$ if $i = 0$ and $[x_{N-1}, b]$ if $i = N$).

The set of hat-functions $\{\varphi_i\}_{i=0}^N$ is a basis for V_h because each $v \in V_h$ has the unique representation

$$v(x) = \sum_{i=0}^{N} v(x_i)\varphi_i(x),$$

where the nodal values $v(x_i)$ appear as coefficients. To see this, it is sufficient to realize that the functions on the left and right hand side are both continuous and piecewise linear and take the same values at the nodes, and thus coincide. Since the number of basis functions φ_i is equal to $N+1$, the dimension of V_h is equal to $N + 1$.

The continuous piecewise linear interpolant $\pi_h f \in V_h$ of a given Lipschitz continuous function $f(x)$ on $[0, 1]$ is defined by

$$\pi_h f(x_i) = f(x_i) \quad \text{for } i = 0, 1, ..., N,$$

that is, $\pi_h f(x)$ interpolates $f(x)$ at the nodes x_i, see Fig. 52.6. We can express $\pi_h f$ in terms of the basis of hat functions $\{\varphi_i\}_{i=0}^N$ as follows:

$$\pi_h f = \sum_{i=0}^{N} f(x_i)\varphi_i \quad \text{or} \quad \pi_h f(x) = \sum_{i=0}^{N} f(x_i)\varphi_i(x) \quad \text{for } x \in [0, 1], \tag{52.12}$$

with the x-dependence indicated.

Since $\pi_h f(x)$ is linear on each subinterval I_i and interpolates $f(x)$ at the end-points of I_i, we can express $\pi f(x)$ analytically on I_i as follows:

$$\pi_h f(x) = f(x_{i-1})\frac{x - x_i}{x_{i-1} - x_i} + f(x_i)\frac{x - x_{i-1}}{x_i - x_{i-1}} \quad \text{for } x_{i-1} \le x \le x_i,$$

for $i = 1, ..., N$.

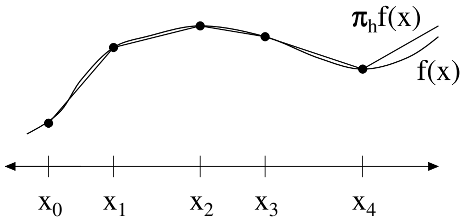


FIGURE 52.6. An example of a continuous piecewise linear interpolant.

Using Theorem 52.2, we obtain the following error estimate for piecewise linear interpolation:

Theorem 52.3 *The piecewise linear interpolant $\pi_h f(x)$ of a twice differentiable function $f(x)$ on a partition of $[a, b]$ with mesh function $h(x)$ satisfies*

$$\begin{aligned} \|f - \pi_h f\|_{L_\infty(a,b)} &\le \frac{1}{8}\|h^2 f''\|_{L_\infty(a,b)}, \\ \|f' - (\pi_h f)'\|_{L_\infty(a,b)} &\le \frac{1}{2}\|h f''\|_{L_\infty(a,b)}. \end{aligned} \tag{52.13}$$

If $f(x)$ is only once differentiable, then

$$\|f - \pi_h f\|_{L_\infty(a,b)} \le \frac{1}{2}\|h f'\|_{L_\infty(a,b)}. \tag{52.14}$$

Note that since the mesh function $h(x)$ may have jumps at the nodes, we interpret $\|h^2 f''\|_{L_\infty(a,b)}$ as

$$\max_{i=1,\dots,N} \max_{y\in[x_{i-1},x_i]} |h^2(y)\, f''(y)|,$$

where $h(y) = x_i - x_{i-1}$ for $y \in [x_{i-1}, x_i]$.

52.5 The L_2 projection into the space V_h of piecewise linear functions on I

Let $f(x)$ be a given function on an interval $I = [a, b]$ and V_h denote the space of continuous piecewise linear functions V_h on a partition $a = x_0 < ... < x_N = b$ of I with mesh function $h(x)$.

The *orthogonal projection* $P_h f$ of the function f into V_h is the function $P_h f \in V_h$ such that

$$\int_I (f - P_h f)v\, dx = 0 \quad \text{for } v \in V_h. \tag{52.15}$$

Recalling the definition of the $L_2(I)$-scalar product

$$(v, w)_{L_2(I)} = \int_I v(x)w(x)\, dx,$$

with the corresponding $L_2(I)$-norm

$$\|v\|_{L_2(I)} = (\int_I v^2(x)\, dx)^{1/2},$$

we can write (52.15) in the form

$$(f - P_h f, v)_{L_2(I)} = 0 \quad \text{for } v \in V_h.$$

This says that $f - P_h f$ is orthogonal to V_h with respect to the $L_2(I)$ scalar product. We also call $P_h f$ the $L_2(I)$*-projection* of f onto V_h.

We first show that $P_h f$ is uniquely defined and then prove that $P_h f$ is the best V_h-approximation of f in the $L_2(I)$-norm.

To prove uniqueness and existence, we express $P_h f$ in the nodal basis $\{\varphi_i\}_{i=0}^N$:

$$P_h f(x) = \sum_{j=0}^{N} c_j \varphi_j(x),$$

where the $c_j = (P_h f)(x_j)$ are the nodal values of $P_h f$ that have to be determined. We insert this representation into (52.15) and choose $v = \varphi_i$ with $i = 0, ..., N$, to get

$$\begin{aligned}\int_I \sum_{j=0}^{N} c_j \varphi_j(x)\varphi_i(x)\, dx &= \sum_{j=0}^{N} c_j \int_I \varphi_j(x)\varphi_i(x)\, dx \\ &= \int_I f\varphi_i\, dx \equiv b_i \quad i = 0, 1, ..., N, \end{aligned} \tag{52.16}$$

where we changed the order of integration and summation. This gives the following system of equations

$$\sum_{j=0}^{N} m_{ij} c_j = \int_I f\varphi_i\, dx \equiv b_i \quad i = 0, 1, ..., N, \tag{52.17}$$

where

$$m_{ij} = \int_I \varphi_j(x)\varphi_i(x)\, dx, \quad i, j = 0, ..., N.$$

We can write (52.17) in matrix form as

$$Mc = b$$

where $c = (c_0, ..., c_N)$ is a $N+1$-vector of the unknown coefficients c_j, and $b = (b_0, ..., b_N)$ is computable from $f(x)$, and $M = (m_{ij})$ is a $(N+1)\times(N+1)$-matrix that depends on the basis functions φ_i, but not on the function $f(x)$. We refer to the matrix M as the *mass matrix.*

We can now easily prove the uniqueness of $P_h f$. Since the difference $P_h f - \bar{P}_h f$ of two functions $P_h f \in V_h$ and $\bar{P}_h f \in V_h$ satisfying the relation (52.15), also satisfy

$$\int_I (P_h f - \bar{P}_h f) v \, dx = 0 \quad \text{for } v \in V_h,$$

by choosing $v = P_h f - \bar{P}_h f$, we get

$$\int_I (P_h f - \bar{P}_h f)^2 \, dx = 0,$$

and thus $P_h f(x) = \bar{P}_h f(x)$ for $x \in I$. Solutions of the system $Mc = b$ are therefore unique, and since M is a square matrix, existence follows from the Fundamental Theorem of Linear Algebra. We sum up:

Theorem 52.4 *The $L_2(I)$-projection $P_h f$ of a given function f onto the set of piecewise linear functions V_h on I is uniquely defined by* (52.15) *or the equivalent system of equations $Mc = b$, where $c_j = P_h f(x_j)$ are the nodal values of $P_h f$, M is the mass matrix with coefficients $m_{ij} = (\varphi_j, \varphi_i)_{L_2(I)} = (\varphi_i, \varphi_j)_{L_2(I)}$ and the coefficients of the right hand side b are given by $b_i = (f, \varphi_i)$.*

EXAMPLE 52.1. We compute the mass matrix M in the case of a uniform subdivision with $h(x) = h = (b-a)/N$ for $x \in I$. We get by a direct computation

$$m_{ii} = \int_{x_{i-1}}^{x_{i+1}} \varphi_i^2(x) \, dx = \frac{2h}{3} \quad i = 1, ...N-1, \qquad m_{00} = m_{NN} = \frac{h}{3},$$

$$m_{i,i+1} = \int_{x_{i-1}}^{x_{i+1}} \varphi_i(x)\varphi_{i+1}(x) \, dx = \frac{h}{6} \quad i = 1, ...N-1.$$

The corresponding "lumped" mass matrix $\hat{M} = (\hat{m}_{ij})$, which is a diagonal matrix with the diagonal element in each row being the sum of the elements in the corresponding row of M, takes the form

$$\hat{m}_{ii} = h \quad i = 1, ..., N-1, \quad \hat{m}_{00} = \hat{m}_{NN} = h/2.$$

We see that $\hat{M}$ may be viewed as a h-scaled variant of the identity matrix and M can be viewed as an h-scaled approximation of the identity matrix.

We now prove that the $L_2(I)$-projection P_hf of a function f satisfies

$$\|f - P_hf\|_{L_2(I)} \leq \|f - v\|_{L_2(I)}, \quad \text{for all } v \in V_h. \tag{52.18}$$

This implies that P_hf is the element in V_h with smallest deviation from f in the $L_2(I)$-norm. Applying Cauchy's inequality to (52.15) with $v \in V_h$ gives

$$\begin{aligned}
&\int_I (f - P_hf)^2\, dx \\
&= \int_I (f - P_hf)(f - P_hf)\, dx + \int_I (f - P_hf)(P_hf - v)\, dx \\
&= \int_I (f - P_hf)(f - v)\, dx \leq \Big(\int_I (f - P_hf)^2\, dx\Big)^{1/2}\Big(\int_I (f - v)^2\, dx\Big)^{1/2},
\end{aligned}$$

which proves the desired result. We summarize:

Theorem 52.5 *The $L_2(I)$-projection P_h into V_h defined by* (52.15)*, is the unique element in V_h which minimizes $\|f - v\|_{L_2(I)}$ with v varying over V_h.*

In particular, choosing $v = \pi_hf$ in (52.18), we obtain

$$\|f - P_hf\|_{L_2(I)} \leq \|f - \pi_hf\|_{L_2(I)},$$

where π_hf is the nodal interpolant of f introduced above. One can prove the following analog of (52.13)

$$\|f - \pi_hf\|_{L_2(I)} \leq \frac{1}{\pi^2}\|h^2 f''\|_{L_2(I)},$$

where the interpolation constant happens to be π^{-2}. We thus conclude the following basic result:

Theorem 52.6 *The $L_2(I)$-projection P_h into the space of piecewise linear functions V_h on I with mesh function $h(x)$, satisfies the following error estimate:*

$$\|f - P_hf\|_{L_2(I)} \leq \frac{1}{\pi^2}\|h^2 f''\|_{L_2(I)}. \tag{52.19}$$

Chapter 52 Problems

52.1. Give a different proof of the first estimate of Theorem Theorem 52.1 by considering for a given $x \in (0,1)$, the function

$$g(y) = f(y) - \pi f(y) - \gamma(x)y(1-y), \quad y \in [0,1],$$

where $\gamma(x)$ is chosen so that $g(x) = 0$. Hint: the function $g(y)$ vanishes at 0, x and 1. Show by repeated use of the Mean Value theorem that g'' vanishes at some point ξ, from which it follows that $\gamma(x) = -f''(\xi)/2$.

52.2. Prove Theorem 52.2 from Theorem 52.1 by using the change of variables $x = a + (b-a)z$ transforming the interval $[0,1]$ onto $[a,b]$, setting $F(z) = f(a + (b-a)z)$ and using that by the Chain Rule, $F' = \frac{dF}{dz} = (b-a)f' = (b-a)\frac{df}{dx}$.

52.3. Develop approximation/interpolation with piecewise constant (discontinuous) functions on a partition of an interval. Consider interpolation at left-hand endpoint, right-hand endpoint, midpoint and mean value for each subinterval. Prove error estimates of the form $\|u - \pi_h u\|_{L_\infty(I)} \leq C\|hu'\|_{L_\infty(I)}$, with $C = 1$ or $C = \frac{1}{2}$.

53
FEM for two-point boundary value problems

> The beginners mind is empty, free of the habits of the expert, ready to accept, or doubt, and open to all the possibilities. It is a kind of mind which can see things as they are. (Shunryu Suzuki)

53.1 Introduction

We begin by deriving a model that is based on a *conservation principle* which states:

> The rate at which a specified quantity changes in a region is equal to the rate that the quantity leaves and enters the region plus the rate at which the quantity is created and destroyed inside the region.

Such a conservation principle holds for a wide variety of quantities, including animals, automobiles, bacteria, chemicals, fluids, heat and energy, etc. So the model we derive has a wide application.

In this chapter, we assume that the quantity to be modeled exists in a very small diameter "tube" with constant cross section and that the quantity varies in the direction along the tube but not at all within a fixed cross section, see Fig. 53.1. We use x to denote the position along the length of the tube and let t denote time. We assume that the quantity in the tube is sufficiently abundant that it makes sense to talk about a *density* $u(x,t)$, measured in amount of the quantity per unit volume, that varies continuously with the position x and time t. This is certainly valid for quantities such as heat and energy, and may be more or less valid for quantities such as bacteria and chemicals provided there is a sufficient number of creatures or molecules respectively.

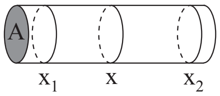


FIGURE 53.1. Variation in a very narrow "tube".

We next express the conservation principle mathematically. We consider a small region of the tube of width dx and cross-sectional area A. The amount of quantity in this region is $u(x,t)A\,dx$. We let $q(x,t)$ denote the *flux* at position x and time t, or the amount of the quantity crossing the section at x at time t measured in amount per unit area per unit time. We choose the orientation so that q is positive when the flow is to the right. The amount of quantity crossing the section at position x at time t is therefore $Aq(x,t)$. Lastly, we let $f(x,t)$ denote the rate that the quantity is created or destroyed within the section at x at time t measured in amount per unit volume per unit time. So, $f(x,t)Adx$ is the amount of the quantity created or destroyed in the small region of width dx per unit time.

The conservation principle for a fixed length of pipe between $x = x_1$ and $x = x_2$ implies that the rate of change of the quantity in this section must equal the rate at which it flows in at $x = x_1$ minus the rate at which it flows out at $x = x_2$ plus the rate at which it is created in $x_1 \leq x \leq x_2$. In mathematical terms,

$$\frac{\partial}{\partial t}\int_{x_1}^{x_2} u(x,t)A\,dx = Aq(x_1,t) - Aq(x_2,t) + \int_{x_1}^{x_2} f(x,t)A\,dx.$$

or

$$\frac{\partial}{\partial t}\int_{x_1}^{x_2} u(x,t)\,dx = q(x_1,t) - q(x_2,t) + \int_{x_1}^{x_2} f(x,t)\,dx. \tag{53.1}$$

Equation (53.1) is called the *integral formulation* of the conservation principle.

We can reformulate (53.1) as a partial differential equation provided $u(x,t)$ and $q(x,t)$ are sufficiently smooth. For we can write,

$$\begin{aligned}\frac{\partial}{\partial t}\int_{x_1}^{x_2} u(x,t)\,dx &= \int_{x_1}^{x_2}\frac{\partial}{\partial t}u(x,t)\,dx,\\ q(x_1,t) - q(x_2,t) &= \int_{x_1}^{x_2}\frac{\partial}{\partial x}q(x,t)\,dx,\end{aligned}$$

and therefore collecting terms,

$$\int_{x_1}^{x_2}\big(\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}q(x,t) - f(x,t)\big)\,dx = 0.$$

Since x_1 and x_2 are arbitrary, the integrand must be zero at each point, or

$$\frac{\partial}{\partial t}u(x,t) + \frac{\partial}{\partial x}q(x,t) = f(x,t). \tag{53.2}$$

Equation (53.2) is the *pointwise* or *differential formulation* of the conservation principle.

So far we have one equation for two unknowns. To complete the model, we use a *constitutive relation* that describes the relation between the flux and the density. This relation is specific to the physical properties of the quantity being modeled, yet it is often unclear exactly how to model these properties. A constitutive relation used in practice is often only an approximation to the true unknown relation.

Many quantities have the property that the quantity flows from regions of high concentration to regions of low concentration, and the rate of flow increases as the differences in concentration increases. As a first approximation, we assume a simple linear relation

$$q(x,t) = -a(x,t)\frac{\partial}{\partial x}u(x,t), \tag{53.3}$$

where $a(x,t) > 0$ is the *diffusion coefficient*. In case u represents heat, (53.3) is known as *Newton's Heat Law*. In general, equation (53.3) is known as *Fick's Law*. Note that the choice of sign of a guarantees for example that flow is to the right if $u_x < 0$, i.e. if u decreases across the section at x. Substituting (53.3) into (53.2), we obtain the general time-dependent reaction-diffusion equation,

$$\frac{\partial}{\partial t}u(x,t) - \frac{\partial}{\partial x}\big(a(x,t)\frac{\partial}{\partial x}u(x,t)\big) = f(x,t).$$

To simplify the notation, we use $\dot{u}$ to denote $\partial u/\partial t$ and u' to denote $\partial u/\partial x$. This yields

$$\dot{u}(x,t) - (a(x,t)u'(x,t))' = f(x,t). \tag{53.4}$$

Convection or transport is another important process to take into account in this model.

EXAMPLE 53.1. When modeling populations of animals, diffusion reflects the natural tendency of most creatures to spread out over a region due to randomly occurring interactions between pairs of creatures, while convection models phenomena such as migration.

Convection is modeled by assuming a constitutive relation in which the flux is proportional to the density, i.e.

$$\varphi(x,t) = b(x,t)u(x,t),$$

which results in a convection term in the differential equation of the form $(bu)'$. The convection coefficient $b(x,t)$ determines the rate and direction of transport of the quantity being modeled.

In general, many quantities are modeled by a constitutive relation of the form

$$\varphi(x,t) = -a(x,t)u'(x,t) + b(x,t)u(x,t)$$

which combines diffusion and convection. Arguing as above, we obtain the general reaction-diffusion-convection equation

$$\dot{u}(x,t) - (a(x,t)u'(x,t))' + (b(x,t)u(x,t))' = f(x,t). \tag{53.5}$$

53.2 Initial boundary-value problems

We have to add suitable data to (53.4) or (53.5) in order to specify a unique solution. We model the amount of substance in a fixed length of tube located between $x = 0$ and $x = 1$, as in Fig. 53.1, and specify some information about u called *boundary conditions* at $x = 0$ and $x = 1$. We also need to give some initial data at some initial time, which we take to be $t = 0$. The *evolutionary* or *time-dependent initial two point boundary value problem* reads: find $u(x,t)$ such that

$$\begin{cases} \dot{u} - (au')' + (bu)' = f & \text{in } (0,1) \times (0,T), \\ u(0,t) = u(1,t) = 0 & \text{for } t \in (0,T) \\ u(x,0) = u_0(x) & \text{for } x \in (0,1), \end{cases} \tag{53.6}$$

where a, b, c are given coefficients and f and g are given data. The boundary values $u(0,t) = u(1,t) = 0$ are known as *homogeneous Dirichlet boundary conditions.*

EXAMPLE 53.2. In the case that we use (53.4) to model the heat u in a long thin wire, the coefficient a represents the heat conductivity of the metal in the wire, f is a given heat source, and the homogeneous Dirichlet boundary conditions at the end-points means that the temperature of the wire is held fixed at 0 there. Such conditions are realistic for example if the wire is attached to very large masses at the ends.

Other boundary conditions found in practice include: *nonhomogeneous Dirichlet boundary conditions* $u(0) = u_0$, $u(1) = u_1$ with constants u_0, u_1; one homogeneous Dirichlet $u(0) = 0$ and one *nonhomogeneous Neumann boundary condition* $a(1)u'(1) = g_1$ with constant g_1; and more general *Robin boundary conditions*

$$-a(0)u'(0) = \gamma(0)(u_0 - u(0)), \quad a(1)u'(1) = \gamma(1)(u_1 - u(1))$$

with constants $\gamma(0)$, u_0, $\gamma(1)$, u_1.

53.3 Stationary boundary value problems

In many situations, u is independent of time and the model reduces to the *stationary* reaction-diffusion equation

$$-(a(x)u'(x))' = f(x) \tag{53.7}$$

in the case of pure diffusion and

$$-(a(x)u'(x))' + (b(x)u(x))' = f(x) \tag{53.8}$$

in case there is convection as well. For these problems, we only need to specify boundary conditions. For example, we consider the *two-point boundary value problem*: find the function $u(x)$ satisfying

$$\begin{cases} -(au')' = f & \text{in } (0,1), \\ u(0) = u(1) = 0 \end{cases} \tag{53.9}$$

and when there is convection: find $u(x)$ such that

$$\begin{cases} -(au')' + (bu)' = f & \text{in } (0,1), \\ u(0,t) = u(1,t) = 0. \end{cases} \tag{53.10}$$

53.4 The Finite Element Method

We begin the discussion of discretization by studying the simplest model above, namely the two-point boundary value problem for the stationary reaction-diffusion model (53.9).

We can express the solution $u(x)$ of (53.9) analytically in terms of data by integrating twice (setting $w = au'$)

$$u(x) = \int_0^x \frac{w(y)}{a(y)}\,dy + \alpha_1, \quad w(y) = -\int_0^y f(z)\,dz + \alpha_2,$$

where the constants α_1 and α_2 are chosen so that $u(0) = u(1) = 0$. We can use this solution formula to compute the value of the solution $u(x)$ for any given $x \in (0,1)$ by evaluating the integrals analytically or numerically using quadrature. However, this is very time consuming if we want the solution at many points in $[0,1]$. This motivates consideration of an alternative way of computing the solution $u(x)$ using the *Finite Element Method* (FEM), which is a general method for solving differential equations numerically. FEM is based on rewriting the differential equation in *variational form* and seeking an approximate solution as a piecewise polynomial.

Note that we do not use the solution by integration outlined above, one important consequence of that procedure is that u is "twice as differentiable" as the data f, since we integrate twice to get from f to u.

We present FEM for (53.9) based on continuous piecewise linear approximation. We let $\mathcal{T}_h : 0 = x_0 < x_1 < ... < x_{M+1} = 1$, be a *partition* (or *triangulation*) of $I = (0,1)$ into sub-intervals $I_j = (x_{j-1}, x_j)$ of length $h_j = x_j - x_{j-1}$. We look for an approximate solution in the set V_h of continuous piecewise linear functions $v(x)$ on $\mathcal{T}_h$ such that $v(0) = 0$ and $v(1) = 0$. We show an example of such a function in Fig. 53.2. In Chapter 52, we saw

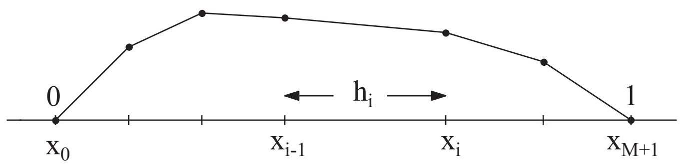


FIGURE 53.2. A continuous piecewise linear function in V_h.

that V_h is a finite dimensional vector space of dimension M with a basis consisting of the hat functions $\{\varphi_j\}_{j=1}^M$ illustrated in Fig. 52.5, associated with the *interior nodes* $x_1, \cdots, x_M$. The coordinates of a function v in V_h in this basis are the values $v(x_j)$ at the interior nodes since a function $v \in V_h$ can be written

$$v(x) = \sum_{j=1}^{M} v(x_j)\varphi_j(x).$$

Note that because $v \in V_h$ is zero at 0 and 1, we do not include φ_0 and φ_{M+1} in the set of basis functions for V_h.

The finite element method is based on restating the differential equation $-(au')' = f$ in an average or *variational* form

$$-\int_0^1 (au')'v\,dx = \int_0^1 fv\,dx, \tag{53.11}$$

where the function v varies over an appropriate set of *test functions*. The variational form results from multiplying the differential equation $-(au')' = f$ by the test function $v(x)$ and integrating over the interval $(0,1)$. The variational formulation says that the residual $-(au')' - f$ of the true solution is orthogonal to all test functions v with respect to the $L_2(0,1)$ scalar product.

The basic idea of FEM is to compute an approximate solution $U \in V_h$ that satisfies (53.11) for a restricted set of test functions. This approach to computing an approximate solution is known as the *Galerkin method* in memory of the Russian engineer and scientist Galerkin (1871-1945), see Fig. 53.3. He invented his method while imprisoned for anti-Tsarist activities during 1906-7. We call the set V_h, where we seek the FEM-solution U, the *trial space* and we call the space of test functions the *test space*. In the present case of homogeneous Dirichlet boundary conditions, we usually choose the test space to be equal to V_h. Consequently, the dimensions of the trial and test spaces are equal, which is necessary for the existence and uniqueness of the approximate solution U.

FIGURE 53.3. Boris Galerkin, inventor of the Finite Element Method

However since the functions in V_h do not have second derivatives, we can not simply plug a potential approximate solution U in V_h directly into (53.11). To get around this difficulty, we use integration by parts to move one derivative from $(au')'$ onto v, noting that functions in V_h are piecewise differentiable. Assuming v is differentiable and $v(0) = v(1) = 0$:

$$\begin{aligned}-\int_0^1 (au')'v\,dx &= -a(1)u'(1)v(1) + a(0)u'(0)v(0) + \int_0^1 au'v'\,dx\\ &= \int_0^1 au'v'\,dx.\end{aligned}$$

This leads to the *continuous Galerkin finite element method of order* 1 (*cG(1)-method*) for (53.9): compute $U \in V_h$ such that

$$\int_0^1 aU'v'\,dx = \int_0^1 fv\,dx \quad \text{for all } v \in V_h. \tag{53.12}$$

We note that the derivatives U' and v' of the functions U and $v \in V_h$ are piecewise constant functions of the form depicted in Fig. 53.4 and are not defined at the nodes x_i. However, the value of an integral is independent of the value of the integrand at isolated points. Therefore, the integral (53.12) with integrand $aU'v'$ is uniquely defined as the sum of the integrals over the sub-intervals I_j.

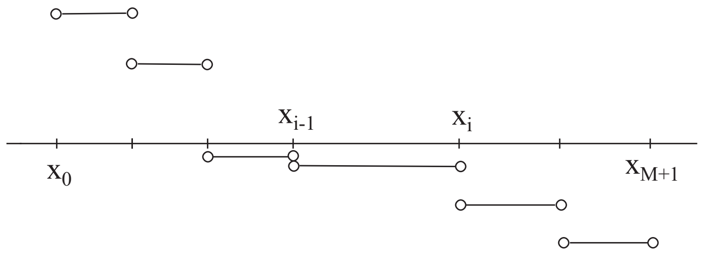


FIGURE 53.4. The derivative of the continuous piecewise linear function in Fig. 53.2.

53.4.1 Discretization of the stationary reaction-diffusion-convection problem

To solve (53.10) numerically let $0 = x_0 < x_1 < ... < x_{L+1} = 1$ be a partition of $(0,1)$, and let V_h be the corresponding space of continuous piecewise linear functions $v(x)$ such that $v(0) = v(1) = 0$. The cG(1) FEM for (53.10) takes the form: compute $U \in V_h$ such that

$$\int_0^1 (aU')v' + (bU)'v)\,dx = \int_0^1 fv\,dx \quad \text{for all } v \in V_h.$$

53.5 The discrete system of equations

We have not yet proved that the set of equations (53.12) has a unique solution nor discussed what is involved in computing the solution U. This is an important issue considering we constructed the FEM precisely because the original problem is likely impossible to solve analytically.

We prove that the cG(1)-method (53.12) corresponds to a square linear system of equations for the unknown nodal values $\xi_j = U(x_j)$, $j = 1, ..., M$. We write U using the basis of hat functions as

$$U(x) = \sum_{j=1}^{M} \xi_j \varphi_j(x) = \sum_{j=1}^{M} U(x_j)\varphi_j(x).$$

Substituting into (53.12), we change the order of summation and integration to obtain

$$\sum_{j=1}^{M} \xi_j \int_0^1 a\varphi_j' v'\,dx = \int_0^1 fv\,dx, \tag{53.13}$$

for all $v \in V_h$. Now, it suffices to check (53.13) with v varying over the set of basis functions $\{\varphi_i\}_{i=1}^M$, since any function in V_h can be expressed as a linear combination of the basis functions. We are thus led to the $M \times M$ linear system of equations

$$\sum_{j=1}^{M} \xi_j \int_0^1 a\varphi_j' \varphi_i' \, dx = \int_0^1 f\varphi_i \, dx, \quad i = 1, ..., M, \tag{53.14}$$

for the unknown coefficients $\xi_1, ..., \xi_M$. We let $\xi = (\xi_1, ..., \xi_M)^\top$ denote the M-vector of unknown coefficients and define the $M \times M$ *stiffness matrix* $A = (a_{ij})$ with elements

$$a_{ij} = \int_0^1 a\varphi_j' \varphi_i' \, dx, \qquad i, j = 1, ..., M,$$

and the *load vector* $b = (b_i)$ with

$$b_i = \int_0^1 f\varphi_i \, dx, \qquad i = 1, ..., M.$$

These names originate from early applications of the finite element method in *structural mechanics* describing deformable structures like the body and wing of an aircraft or buildings. Using this notation, (53.14) is equivalent to the system of linear equations

$$A\xi = b. \tag{53.15}$$

In order to solve for the unknown vector ξ of nodal values of U, we first have to compute the stiffness matrix A and the load vector b. In the first instance, we assume that $a(x) = 1$ for $x \in [0, 1]$. We note that a_{ij} is zero unless $i = j - 1$, $i = j$, or $i = j + 1$ because otherwise either $\varphi_i(x)$ or $\varphi_j(x)$ is zero on each sub-interval occurring in the integration. We illustrate this in Fig. 53.5. We compute a_{ii} first. Using the definition of the hat function

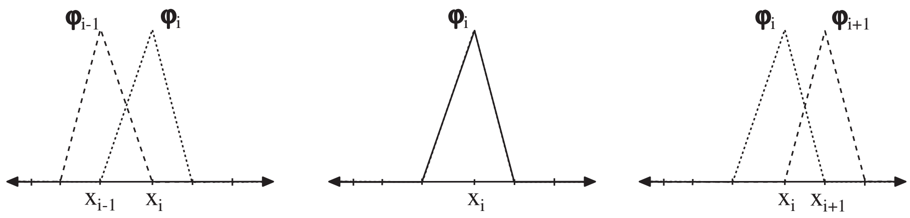


FIGURE 53.5. Three possibilities to obtain a non-zero element in the stiffness matrix.

φ_i,

$$\varphi_i(x) = \begin{cases} (x - x_{i-1})/h_i, & x_{i-1} \le x \le x_i, \\ (x_{i+1} - x)/h_{i+1}, & x_i \le x \le x_{i+1}, \\ 0, & \text{elsewhere}, \end{cases}$$

the integration breaks down into two integrals:

$$a_{ii} = \int_{x_{i-1}}^{x_i} \big(\frac{1}{h_i}\big)^2 \, dx + \int_{x_i}^{x_{i+1}} \big(\frac{-1}{h_{i+1}}\big)^2 \, dx = \frac{1}{h_i} + \frac{1}{h_{i+1}} \text{ for } i = 1, 2, ..., M,$$

since $\varphi_i' = 1/h_i$ on (x_{i-1}, x_i) and $\varphi_i' = -1/h_{i+1}$ on (x_i, x_{i+1}) and φ_i is zero on the other sub-intervals. Similarly,

$$a_{i\,i+1} = \int_{x_i}^{x_{i+1}} \frac{-1}{(h_{i+1})^2} \, dx = -\frac{1}{h_{i+1}} \text{ for } i = 1, 2, ..., M,$$

while $a_{i\,i-1} = -1/h_i$ for $i = 2, 3, ..., M$.

We compute the elements of the load vector of b in the same way to get

$$b_i = \int_{x_{i-1}}^{x_i} f(x) \frac{x - x_{i-1}}{h_i} dx + \int_{x_i}^{x_{i+1}} f(x) \frac{x_{i+1} - x}{h_{i+1}} dx, \quad i = 1, ..., M.$$

The matrix A is a *sparse* matrix in the sense that most of its entries are zero. In particular, A is a *banded* matrix with non-zero entries occurring only in the diagonal, super-diagonal and sub-diagonal positions. A is also called a tri-diagonal matrix. Moreover, A is a *symmetric* matrix since $\int_0^1 \varphi_i' \varphi_j' \, dx = \int_0^1 \varphi_j' \varphi_i' \, dx$. Finally, A is *positive-definite* in the sense that

$$\eta^\top A \eta = \sum_{i,j=1}^{M} \eta_i a_{ij} \eta_j > 0,$$

unless $\eta_i = 0$ for $i = 1, ..., M$. This follows by noting that if $v(x) = \sum_{j=1}^{M} \eta_j \varphi_j(x)$ then by reordering the summation (check!)

$$\begin{aligned} \sum_{i,j=1}^{M} \eta_i a_{ij} \eta_j &= \sum_{i,j=1}^{M} \eta_i \int_0^1 a \varphi_j' \varphi_i' \, dx \, \eta_j \\ &= \int_0^1 a \sum_{j=1}^{M} \eta_j \varphi_j' \sum_{i=1}^{M} \eta_i \varphi_i' \, dx = \int_0^1 a v'(x) v'(x) \, dx > 0 \end{aligned}$$

unless $v'(x) = 0$ for all $x \in [0, 1]$, that is $v(x) = 0$ for $x \in [0, 1]$, since $v(0) = 0$, that is $\eta_i = 0$ for $i = 1, ..., M$. This implies that A is invertible, so that (53.15) has a unique solution for all data b.

We sum up: the stiffness matrix A is sparse, symmetric and positive definite, and thus in particular the system $A\xi = b$ has a unique solution for all b.

We expect the accuracy of the approximate solution to increase as M increases since the work involved in solving for U increases. Systems of dimension $10^2 - 10^3$ in one space dimension and up to 10^6 in two or three space dimensions are common. An important issue is the efficient numerical solution of the system $A\xi = b$.

53.6 Discretizing problems with different boundary conditions

We consider briefly the discretization of the two-point boundary value problem $-(au')' = f$ in $(0,1)$ with the different boundary conditions.

53.6.1 Non-homogeneous Dirichlet boundary conditions

We begin with the boundary conditions $u(0) = u_0$ and $u(1) = u_1$, where u_0 and u_1 are given boundary values, where the conditions are non-homogeneous if $u_0 u_1 \neq 0$. In this situation, we compute an approximate solution in the trial space V_h of continuous piecewise linear functions $v(x)$ on a partition $\mathcal{T}_h : 0 = x_0 < x_1 < ... < x_{M+1} = 1$, satisfying the boundary conditions $v(0) = u_0$, $v(1) = u_1$, and we let the test functions vary over the space V_h^0 of continuous piecewise linear functions $v(x)$ satisfying the homogeneous boundary conditions $v(0) = v(1) = 0$. The trial and test spaces are different in this case, but we note that they have equal dimension (equal to the number M of internal nodes). Multiplying by a test function and integrating by parts, we are led to the following method: compute $U \in V_h$ such that

$$\int_0^1 aU'v'\,dx = \int_0^1 fv\,dx \quad \text{for all } v \in V_h^0. \tag{53.16}$$

As above this leads to a symmetric positive definite system of equations in the internal unknown nodal values $U(x_1), ..., U(x_M)$.

53.6.2 Neumann boundary conditions

We now consider the problem

$$\begin{cases} -(au')' = f, & \text{in } (0,1), \\ u(0) = 0, \; a(1)u'(1) = g_1, \end{cases} \tag{53.17}$$

with a non-homogeneous Neumann boundary condition at $x = 1$, which in the case of modeling heat in a wire, corresponds to prescribing the heat flux $a(1)u'(1)$ at $x = 1$ to be g_1.

To derive a variational formulation of this problem, we multiply the differential equation $-(au')' = f$ by a test function v and integrate by parts to get

$$\int_0^1 fv\,dx = -\int_0^1 (au')'v\,dx = \int_0^1 au'v'\,dx - a(1)u'(1)v(1) + a(0)u'(0)v(0).$$

Now $a(1)u'(1) = g_1$ is specified but $a(0)u'(0)$ is unknown. So it is convenient to assume that v satisfies the homogeneous Dirichlet condition $v(0) = 0$.

Correspondingly, we define V_h to be the space of continuous functions v that are piecewise linear on a partition $\mathcal{T}_h$ of $(0,1)$ satisfying $v(0) = 0$. Replacing $a(1)u'(1)$ by g_1, we are led to the following FEM for (53.17): compute $U \in V_h$ such that

$$\int_0^1 aU'v'\,dx = \int_0^1 fv\,dx + g_1 v(1) \quad \text{for all } v \in V_h. \tag{53.18}$$

We substitute $U(x) = \sum_{i=1}^{M+1} \xi_i \varphi_i(x)$, noting that the value $\xi_{M+1} = U(x_{M+1})$ at the node x_{M+1} is now undetermined, into (53.18) and choose $v = \varphi_1, \cdots, \varphi_{M+1}$ to get a $(M+1) \times (M+1)$ system of equations for ξ. We show the form of the resulting stiffness matrix with $a = 1$ and load vector in Fig. 53.6. Note that the last equation

$$\left(\begin{array}{c|c} \begin{matrix} \\ \\ A \\ \\ \\ \end{matrix} & \begin{matrix} 0 \\ \vdots \\ 0 \\ -h_{M+1}^{-1} \end{matrix} \\ \hline 0 \;\cdots\; 0 \;\; -h_{M+1}^{-1} & h_{M+1}^{-1} \end{array}\right) \qquad \left(\begin{array}{c} \begin{matrix} \\ \\ b \\ \\ \\ \end{matrix} \\ \hline b_{M+1}+g_1 \end{array}\right)$$

FIGURE 53.6. The stiffness matrix and load vector computed from (53.18) in the case that $a \equiv 1$. A and b are the stiffness matrix and load vector previously obtained in the problem with homogeneous Dirichlet boundary conditions and $b_{M+1} = \int_0^1 f\varphi_{M+1}\,dx$.

$$\frac{U(x_{M+1}) - U(x_M)}{h_{M+1}} = b_{M+1} + g_1$$

is a discrete analog of the boundary condition $u'(1) = g_1$ since $b_{M+1} \approx \frac{h_{M+1}}{2} f(1)$.

To conclude, a Neumann boundary condition, unlike a Dirichlet condition, is not explicitly enforced in the trial space. Instead, the Neumann condition is automatically satisfied as a consequence of the variational formulation by letting the test functions vary freely at the corresponding boundary point. In the case of Neumann boundary conditions, we thus simply can "forget" the boundary conditions in the definition of the trial space V_h and let the test space coincide with V_h. A Dirichlet boundary condition is called an *essential* boundary condition and a Neumann condition is called a *natural* boundary condition. An essential boundary condition is imposed explicitly in the definition of the trial space, i.e. it is a *strongly imposed* boundary condition, and the test space satisfy the corresponding homogeneous boundary condition. A natural boundary condition is not imposed in the trial space and becomes automatically satisfied through the variational formulation by letting the test functions vary freely at the corresponding boundary point.

53.6.3 Robin boundary conditions

A natural generalization of Neumann conditions for the problem $-(au')' = f$ in $(0,1)$ are called *Robin* boundary conditions. These take the form

$$-a(0)u'(0) = \gamma(0)(u_0 - u(0)), \quad a(1)u'(1) = \gamma(1)(u_1 - u(1)). \qquad (53.19)$$

In the case of modeling heat in a wire, $\gamma(0)$ and $\gamma(1)$ are given (non-negative) boundary heat conductivities and u_0 and u_1 are given "outside temperatures". The Robin boundary condition at $x = 0$ states that the heat flux $-a(0)u'(0)$ is proportional to the temperature difference $u_0 - u(0)$ between the outside and inside temperature. If $u_0 > u(0)$ then heat will flow from outside to inside and if $u_0 < u(0)$ then heat will flow from inside out.

EXAMPLE 53.3. We may experience this kind of boundary condition with $\gamma(0)$ quite large in a poorly insulated house on a cold winter day. The size of the boundary heat conductivity γ is an important issue in the real estate business in the north of Sweden.

When $\gamma = 0$, (53.19) reduces to a homogeneous Neumann boundary condition. Conversely, letting γ tend to infinity, the Robin boundary condition $-a(0)u'(0) = \gamma(0)(u_0 - u(0))$ approaches the Dirichlet boundary condition $u(0) = u_0$.

Robin boundary conditions are natural boundary conditions like Neumann conditions, Therefore, we let V_h be the space of continuous piecewise linear functions on a partition of $(0,1)$ without any boundary conditions imposed. Multiplying the equation $-(au')' = f$ by a function $v \in V_h$ and integrating by parts, we get

$$\int_0^1 fv\,dx = -\int_0^1 (au')'v\,dx = \int_0^1 au'v'\,dx - a(1)u'(1)v(1) + a(0)u'(0)v(0).$$

Replacing $a(0)u'(0)$ and $a(1)u'(1)$ using the Robin boundary conditions, we get

$$\int_0^1 fv\,dx = \int_0^1 au'v'\,dx + \gamma(1)(u(1) - u_1)v(1) + \gamma(0)(u(0) - u_0)v(0).$$

Collecting data on the right hand side, we are led to the following cG(1) method: compute $U \in V_h$ such that

$$\int_0^1 aU'v'\,dx + \gamma(0)u(0)v(0) + \gamma(1)u(1)v(1) = \int_0^1 fv\,dx + \gamma(0)u_0v(0) + \gamma(1)u_1v(1)$$

for all $v \in V_h$.

An even more general Robin boundary condition has the form $-a(0)u'(0) = \gamma(0)(u_0 - u(0)) + g_0$, where g_0 is a given heat flux. This Robin boundary condition thus includes Neumann boundary conditions ($\gamma = 0$) and Dirichlet boundary conditions (letting $\gamma \to \infty$). The implementation of a Robin boundary conditions is facilitated by the fact that the trial and test space are the same.

53.7 Error estimates and adaptive error control

When conducting scientific experiments in a laboratory or building a suspension bridge, for example, there is always a lot of worry about the errors in the process. In fact, if we were to summarize the philosophy behind the scientific revolution, a main component would be the modern emphasis on the quantitative analysis of error in measurements during experiments and the reporting of the errors along with the results. The same issue comes up in computational mathematical modeling: whenever we make a computation on a practical problem, we must be concerned with the accuracy of the results and the related issue of how to compute efficiently. These issues naturally fit into a wider framework which also addresses how well the differential equation models the underlying physical situation and what effect errors in data and the model have on the conclusions we can draw from the results.

We address these issues by deriving two kinds of error estimates for the error $u - U$ of the finite element approximation. First we prove an *a priori* error estimate which shows that the Galerkin finite element method for (53.9) produces the best possible approximation in V_h of the solution u in a certain sense. If u has continuous second derivatives, then we know that V_h contains good approximations of u, for example the piecewise linear interpolant. So the a priori estimate implies that the error of the finite element approximation can be made arbitrarily small by refining the mesh provided that the solution u is sufficiently smooth to allow the interpolation error to go to zero as the mesh is refined. This kind of result is called an *a priori* error estimate because the error bound does not depend on the approximate solution to be computed. One the other hand, it does requires knowledge about the derivatives of the (unknown) exact solution.

After that, we prove an *a posteriori* error bound that bounds the error of the finite element approximation in terms of its residual error. This error bound can be evaluated once the finite element solution has been computed and used to estimate the error. Through the a posteriori error estimate, it is possible to estimate and adaptively control the finite element error to a desired tolerance level by suitably refining the mesh.

To measure the size of the error $e = u - U$, we shall use the *weighted L_2 norm*

$$\|w\|_a = \left(\int_0^1 a\, w^2\, dx\right)^{1/2},$$

with *weight a*. More precisely we shall estimate the quantity

$$\|(u-U)'\|_a$$

which we refer to as the *energy norm* of the error $u - U$.

We will use the following variations of Cauchy's inequality with the weight a present:

$$\left|\int_0^1 av'w'\, dx\right| \leq \|v'\|_a\|w'\|_a \quad \text{and} \quad \left|\int_0^1 vw\, dx\right| \leq \|v\|_a\|w\|_{a^{-1}}. \tag{53.20}$$

53.7.1 An a priori error estimate

We shall prove that the finite element approximation $U \in V_h$ is the best approximation of u in V_h with respect to the energy norm. This is a consequence of the *Galerkin orthogonality* built into the finite element method expressed by

$$\int_0^1 a(u-U)'v'\, dx = 0 \quad \text{for all } v \in V_h \tag{53.21}$$

which results from subtracting (53.12) from (53.11) (integrated by parts) with $v \in V_h$. This is analogous to the best approximation property of the L_2 projection studied in the Chapter Piecewise linear approximation.

We have for any $v \in V_h$,

$$\begin{aligned}
\|(u-U)'\|_a^2 &= \int_0^1 a(u-U)'(u-U)'\, dx \\
&= \int_0^1 a(u-U)'(u-v)'\, dx + \int_0^1 a(u-U)'(v-U)'\, dx \\
&= \int_0^1 a(u-U)'(u-v)'\, dx,
\end{aligned}$$

where the last line follows because $v - U \in V_h$. Estimating using Cauchy's inequality, we get

$$\|(u-U)'\|_a^2 \leq \|(u-U)'\|_a\|(u-v)'\|_a,$$

so that

$$\|(u-U)'\|_a \le \|(u-v)'\|_a \quad \text{for all } v \in V_h.$$

This is the best approximation property of U. We now choose in particular $v = \pi_h u$, where $\pi_h u \in V_h$ is the nodal interpolant of u, and use the following weighted analog of (52.11)

$$\|(u-\pi_h u)'\|_a \le C_i \|hu''\|_a,$$

where C_i is an interpolation constant that depends only on (the variation of) a. We then obtain the following error estimate.

Theorem 53.1 *The finite element approximation U satisfies $\|(u-U)'\|_a \le \|(u-v)'\|_a$ for all $v \in V_h$. In particular, there is a constant C_i depending only on a such that*

$$\|u' - U'\|_a \le C_i \|hu''\|_a.$$

This energy norm estimate says that the derivative of the error of the finite element approximation converges to zero at a first order rate in the mesh size h. By integration it follows that the error itself, say pointwise or in the L_2 norm, aslo tends to zero. One can also prove a more precise bound for the error $u - U$ itself that is second order in the mesh size h.

53.7.2 An a posteriori error estimate

We shall now estimate the energy norm error $\|u' - U'\|_a$ in terms of the residual $R(U) = (aU')' + f$ of the finite element solution U on each subinterval. The residual measures how well U solves the differential equation and it is completely computable once U has been computed.

We start by using the variational form of (53.11) with $v = e = u - U$ to find an expression for $\|u - U\|_a^2$:

$$\begin{aligned} \|e'\|_a^2 &= \int_0^1 ae'e'\,dx = \int_0^1 au'e'\,dx - \int_0^1 aU'e'\,dx \\ &= \int_0^1 fe\,dx - \int_0^1 aU'e'\,dx. \end{aligned}$$

We then use (53.12), with $v = \pi_h e$ denoting the nodal interpolant of e in V_h, to obtain

$$\begin{aligned} \|e'\|_a^2 &= \int_0^1 f\,(e-\pi_h e)\,dx - \int_0^1 aU'(e-\pi_h e)'\,dx \\ &= \int_0^1 f\,(e-\pi_h e)\,dx - \sum_{j=1}^{M+1} \int_{I_j} aU'(e-\pi_h e)'\,dx. \end{aligned}$$

Now, we integrate by parts over each sub-interval I_j in the last term and use the fact that all the boundary terms disappear because $(e - \pi_h e)(x_j) = 0$ to get the *error representation formula*

$$\|e'\|_a^2 = \int_0^1 R(U)(e - \pi_h e)\, dx, \tag{53.22}$$

where the residual error $R(U)$ is the discontinuous function defined on $(0, 1)$ by

$$R(U) = f + (aU')' \quad \text{on each sub-interval } I_j.$$

From the weighted Cauchy inequality (53.20) (inserting factors h and h^{-1}), we get

$$\|e'\|_a^2 \leq \|hR(U)\|_{a^{-1}} \|h^{-1}(e - \pi_h e)\|_a.$$

One can prove the following analog of the second estimate of (52.11)

$$\|h^{-1}(e - \pi_h e)\|_a \leq C_i \|e'\|_a,$$

where C_i is an interpolation constant depending on a, and we notice the appearance of the factor h^{-1} on the left hand side. This proves the basic a posteriori error estimate:

Theorem 53.2 *There is an interpolation constant C_i depending only on a such that the finite element approximation U satisfies*

$$\|u' - U'\|_a \leq C_i \|hR(U)\|_{a^{-1}}. \tag{53.23}$$

53.7.3 Adaptive error control

Since the a posteriori error estimate (53.23) indicates the size of the error of an approximation on a given mesh in terms of computable information, it is natural to try to use this information to compute an accurate approximation. This is the basis of *adaptive error control.*

The computational problem that arises once a two-point boundary value problem is specified is to find a mesh such that the finite element approximation achieves a given level of accuracy, or in other words, such that the error of the approximation is bounded by an *error tolerance* TOL. In practice, we are also concerned with efficiency, which means that we want to determine a mesh with the fewest number of elements that yields an approximation with the desired accuracy. We try to reach this optimal mesh by starting with a coarse mesh and successively refining based on the size of the a posteriori error estimate. By starting with a coarse mesh, we try to keep the number of elements as small as possible.

More precisely, we choose an initial mesh $\mathcal{T}_h$, compute the corresponding cG(1) approximation U, and then check whether or not

$$C_i \|hR(U)\|_{a^{-1}} \leq \text{TOL}.$$

This is the *stopping criterion*, which guarantees that $\|u' - U'\|_a \leq \text{TOL}$ by (53.23). Therefore when the stopping criterion is satisfied, U is sufficiently accurate. If the stopping criterion is not satisfied, we try to construct a new mesh $\mathcal{T}_{\tilde{h}}$ of mesh size $\tilde{h}$ with as few elements as possible such that

$$C_i \|\tilde{h} R(U)\|_{a^{-1}} = \text{TOL}.$$

This is the *mesh modification criterion* from which the new mesh size $\tilde{h}$ is computed based on the size of the residual error $R(U)$ of the approximation on the old mesh. In order to minimize the number of mesh points, it turns out that the mesh size should be chosen to *equidistribute* the residual error in the sense that the contribution from each element to the integral giving the total residual error is roughly the same. In practice, this means that elements with large residual errors are refined, while elements in intervals where the residual error is small are combined together to form bigger elements.

We repeat the adaptive cycle of mesh modification followed by solution on the new mesh until the stopping criterion is satisfied. By the a priori error estimate, we know that if u'' is bounded then the error tends to zero as the mesh is refined. Hence, the stopping criterion will be satisfied eventually. In practice, the adaptive error control rarely requires more than a few iterations.

53.8 Discretization of time-dependent reaction-diffusion-convection problems

We now return to original time dependent problem (53.6).

To solve (53.6) numerically, we apply the cG(1) method for time discretization and the cG(1) FEM for discretization in space. More precisely, let $0 = x_0 < x_1 < ... < x_{L+1} = 1$ be a partition of $(0, 1)$, and let V_h be the corresponding space of continuous piecewise linear functions $v(x)$ such that $v(0) = v(1) = 0$. Let $0 = t_0 < t_1 < t_2 < ... < t_N = T$ be a sequence of discrete time levels with corresponding time intervals $I_n = (t_{n-1}, t_n)$ and time steps $k_n = t_n - t_{n-1}$, for $n = 1, ..., N$. We look for a numerical solution $U(x, t)$ that is linear in t on each time interval I_n. For $n = 1, ..., N$, we compute $U^n \in V_h$ such that for all $v \in V_h$,

$$\begin{aligned} &\int_{I_n} \int_0^1 \dot{U} v \, dx \, dt + \int_{I_n} \int_0^1 (aU')v' + (bU)'v) \, dx \, dt \\ &= \int_{I_n} \int_0^1 f v \, dx \, dt + \int_{I_n} (g(0,t)v(0) + g(1,t)v(1)) \, dt, \end{aligned} \tag{53.24}$$

where $U(t_n, x) = U^n(x)$ denotes the *time nodal value* for $n = 1, 2, ..., N$ and $U^0 = u_0$, assuming that $u_0 \in V_h$. Since U is linear on each time interval, it is determined completely once we have computed its nodal values.

Arguing as above using the expansion in terms of the basis functions for V_h leads to a sequence of systems of equations for $n = 1, ..., N$,

$$MU^n + k_n A_n U^n = MU^{n-1} + k_n b^n, \tag{53.25}$$

where M is the mass matrix corresponding to V_h and A_n is a stiffness matrix related to time interval I_n. Solving this system successively for $n = 1, 2, ..., N$, we obtain an approximate solution U of (53.10).

53.9 Non-linear reaction-diffusion-convection problems

In many situations, the coefficients or data depend on the solution u, which leads to a nonlinear problem. For example if f depends on u, we get a problem of the form

$$\begin{cases} \dot{u} - (au')' + (bu)' = f(u) & \text{in } (0,1) \times (0,T), \\ u(0,t) = u(1,t) = 0, & \text{for } t \in (0,T), \\ u(x,0) = u_0(x) & \text{for } x \in (0,1). \end{cases} \tag{53.26}$$

Discretization as above eventually yields a discrete system of the form

$$MU^n + k_n A_n U^n = MU^{n-1} + k_n b^n(U^n), \tag{53.27}$$

where b^n depends on U^n. This nonlinear system may be solved by fixed point iteration or Newton's method.

We conclude this section by presenting some examples of systems of non-linear reaction-diffusion-convection problems arising in physics, chemistry and biology. These systems may be solved numerically by a direct extension of the cG(1) method in space and time presented above. In all examples, a and the α_i are a positive constants.

EXAMPLE 53.4. *The bistable equation for ferro-magnetism*

$$\dot{u} - au'' = u - u^3, \tag{53.28}$$

with a small.

EXAMPLE 53.5. *Model of a superconductivity of a fluid*

$$\begin{aligned} \dot{u}_1 - au_1'' &= (1 - |u|^2)u_1, \\ \dot{u}_2 - au_2'' &= (1 - |u|^2)u_2. \end{aligned} \tag{53.29}$$

EXAMPLE 53.6. *Model of flame propagation*

$$\begin{aligned} \dot{u}_1 - au_1'' &= -u_1 e^{-\alpha_1/u_2}, \\ \dot{u}_2 - au_2'' &= \alpha_2 u_1 e^{-\alpha_1/u_2}. \end{aligned} \tag{53.30}$$

EXAMPLE 53.7. *Field-Noyes equations for chemical reactions*

$$\begin{aligned}\dot{u}_1 - au_1'' &= \alpha_1(u_2 - u_1u_3 + u_1 - \alpha_2 u_1^2),\\ \dot{u}_2 - au_2'' &= \alpha^{-1}(\alpha_3 u_3 - u_2 - u_1u_2),\\ \dot{u}_2 - au_2'' &= \alpha_4(u_1 - u_3).\end{aligned} \tag{53.31}$$

EXAMPLE 53.8. *Spread of rabies in foxes*

$$\begin{aligned}\dot{u}_1 - au_1'' &= \alpha_1(1 - u_1 - u_2 - u_3) - u_3u_1,\\ \dot{u}_2 - au_2'' &= u_3u_1 - (\alpha_2 + \alpha_3 + \alpha_1u_1 + \alpha_1u_1 + \alpha_1u_3)u_2,\\ \dot{u}_2 - au_2'' &= \alpha_2u_2 - (\alpha_4 + \alpha_1u_1 + \alpha_1u_1 + \alpha_1u_3)u_3,\end{aligned} \tag{53.32}$$

where $\alpha_4 < (1 + (\alpha_3 + \alpha_1)/\alpha_2)^{-1} - \alpha_1$.

EXAMPLE 53.9. *Interaction of two species*

$$\begin{aligned}\dot{u}_1 - au_1'' &= u_1M(u_1, u_2),\\ \dot{u}_2 - au_2'' &= u_2N(u_1, u_2),\end{aligned} \tag{53.33}$$

where $M(u_1, u_2)$ and $N(u_1, u_2)$ are given functions describing various situations such as (i) predator-prey ($M_{u_2} < 0$, $N_{u_1} > 0$) (ii) competing species ($M_{u_2} < 0$, $N_{u_1} < 0$) and (iii) symbiosis ($M_{u_2} > 0$, $N_{u_1} > 0$).

EXAMPLE 53.10. *Morphogenesis of patterns (zebra or tiger)*

$$\begin{aligned}\dot{u}_1 - au_1'' &= -u_1u_2^2 + \alpha_1(1 - u_1)\\ \dot{u}_2 - au_2'' &= u_1u_2^2 - (\alpha_1 + \alpha_2)u_2.\end{aligned} \tag{53.34}$$

EXAMPLE 53.11. *Fitz-Hugh-Nagumo model for transmission of axons*

$$\begin{aligned}\dot{u}_1 - au_1'' &= -u_1(u_1 - \alpha_1)(u_1 - 1) - u_2\\ \dot{u}_2 - au_2'' &= \alpha_2u_1 - \alpha_3u_2,\end{aligned} \tag{53.35}$$

$0 < \alpha_1 < 1$.

Chapter 53 Problems

53.1. Compute the stiffness matrix and load vector for the cG(1) method on a uniform partition for (53.9) with $a(x) = 1 + x$ and $f(x) = \sin(x)$. Use quadrature if exact integration is inconvenient.

53.2. Formulate the cG(1) method for the problem $-(au')' + cu = f$ in $(0, 1)$, $u(0) = u(1) = 0$, where $a(x)$ and $c(x)$ are positive coefficients. Compute the corresponding stiffness matrix when $a = c = 1$, assuming a uniform partition. Is the stiffness matrix still symmetric, positive-definite, and tridiagonal?

53.3. Determine the resulting system of equations corresponding to the cG(1) method (53.16) with non-homogeneous Dirichlet boundary conditions.

53.4. Prove a priori and a posteriori error estimates for cG(1) for $-(au')' = f$ in $(0,1)$ with Robin boundary conditions (a positive).

53.5. Prove a priori and a posteriori error estimates for cG(1) for $-(au')'+cu = f$ in $(0,1)$ with Robin boundary conditions (a and c positive).

> The "classical" phase of my career was summed up in the book *The Large Scale Structure of Spacetime* which Ellis and I wrote in 1973. I would not advise readers of this book to consult that work for further information: it is highly technical, and quite unreadable. I hope that since then I have learned how to write in a manner that is easier to understand. (Stephen Hawking in *A Brief History of Time*)

Part II

Calculus in several dimensions

$$\Delta u = \nabla \cdot \nabla u$$

$$\frac{\partial u}{\partial t} - \Delta u = f$$

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = f$$

$$i\frac{\partial u}{\partial t} - \Delta u + Vu = 0$$

$$\frac{\partial u}{\partial t} + u \cdot \nabla u - \nu \Delta u = f$$

$$\nabla \times H = J, \quad \nabla \cdot B = 0, \quad B = \mu H.$$

54
Vector-valued functions of several real variables

Auch die Chemiker müssen sich allmänlich an den Gedanken gewöhnen, dass ihnen die theoretische Chemie ohne die Beherrschung der Elemente der höheren Analysis ein Buch mit sieben Siegeln blieben wirt. Ein Differential- oder Integralzeichen muss aufhören, für den Chemiker eine unvertständliche Hieroglyphe zu sein,... wenn er sich nicht der Gefahr aussetzen will, für die Entwicklung der theoretischen Chemie jedes Verständnis zu verlieren. (H. Jahn, Grundriss der Elektrochemie, 1895)

54.1 Introduction

We now turn to the extension of the basic concepts of real-valued functions of one real variable, such as Lipschitz continuity and differentiability, to vector-valued functions of several variables. We have carefully prepared the material so that this extension will be as natural and smooth as possible. We shall see that the proofs of the basic theorems like the Chain rule, the Mean Value theorem, Taylor's theorem, the Contraction Mapping theorem and the Inverse Function theorem, extend almost word by word to the more complicated situation of vector valued functions of several real variables.

We consider functions $f : \mathbb{R}^n \to \mathbb{R}^m$ that are vector valued in the sense that the value $f(x) = (f_1(x), ..., f_m(x))$ is a vector in $\mathbb{R}^m$ with components $f_i : \mathbb{R}^n \to \mathbb{R}$ for $i = 1, ..., m$, where with $f_i(x) = f_i(x_1, ..., x_n)$ and $x = (x_1, ..., x_n) \in \mathbb{R}^n$. As usual, we view $x = (x_1, ..., x_n)$ as a n-column vector and $f(x) = (f_1(x), ..., f_m(x))$ as a m-column vector.

As particular examples of vector-valued functions, we first consider *curves*, which are functions $g : \mathbb{R} \to \mathbb{R}^n$, and *surfaces*, which are functions $g : \mathbb{R}^2 \to \mathbb{R}^n$. We then discuss composite functions $f \circ g : \mathbb{R} \to \mathbb{R}^m$, where $g : \mathbb{R} \to \mathbb{R}^n$ is a curve and $f : \mathbb{R}^n \to \mathbb{R}^m$, with $f \circ g$ again being a curve. We recall that $f \circ g(t) = f(g(t))$.

The inputs to the functions reside in the n dimensional vector space $\mathbb{R}^n$ and it is worthwhile to consider the properties of $\mathbb{R}^n$. Of particular importance is the notion of Cauchy sequence and convergence for sequences $\{x^{(j)}\}_{j=1}^{\infty}$ of vectors $x^{(j)} = (x_1^{(j)}, ..., x_n^{(j)}) \in \mathbb{R}^n$ with coordinates $x_k^{(j)}$, $k = 1, ..., n$. We say that the sequence $\{x^{(j)}\}_{j=1}^{\infty}$ is a *Cauchy sequence* if for all $\epsilon > 0$, there is a natural number N so that

$$\|x^{(i)} - x^{(j)}\| \leq \epsilon \quad \text{for } i, j > N.$$

Here $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^n$, that is, $\|x\| = (\sum_{i=1}^n x_i^2)^{1/2}$. Sometimes, it is convenient to work with the norms $\|x\|_1 = \sum_{i=1}^n |x_i|$ or $\|x\|_\infty = \max_{i=1,\dots,n} |x_i|$. We say that the sequence $\{x^{(j)}\}_{j=1}^{\infty}$ of vectors in $\mathbb{R}^n$ *converges* to $x \in \mathbb{R}^n$ if for all $\epsilon > 0$, there is a natural number N so that

$$\|x - x^{(i)}\| \leq \epsilon \quad \text{for } i > N.$$

It is easy to show that a convergent sequence is a Cauchy sequence and conversely that a Cauchy sequence converges. We obtain these results applying the corresponding results for sequences in $\mathbb{R}$ to each of the coordinates of the vectors in $\mathbb{R}^n$.

EXAMPLE 54.1. The sequence $\{x^{(i)}\}_{i=1}^{\infty}$ in $\mathbb{R}^2$ with $x^{(i)} = (1 - i^{-2}, \exp(-i))$, converges to $(1, 0)$.

54.2 Curves in $\mathbb{R}^n$

A function $g : I \to \mathbb{R}^n$, where $I = [a, b]$ is an interval of real numbers, is a *curve* in $\mathbb{R}^n$, see Fig. 54.1. If we use t as the independent variable ranging over I, then we say that the curve $g(t)$ is *parametrized* by the variable t. We also refer to the set of points $\Gamma = \{g(t) \in \mathbb{R}^n : t \in I\}$ as the curve Γ parameterized by the function $g : I \to \mathbb{R}^n$.

EXAMPLE 54.2. The simplest example of a curve is a straight line. The function $g : \mathbb{R} \to \mathbb{R}^2$ given by

$$g(t) = \bar{x} + tz,$$

where $z \in \mathbb{R}^2$ and $\bar{x} \in \mathbb{R}^2$, is a straight line in $\mathbb{R}^2$ through the point $\bar{x}$ with direction z, see Fig. 54.2.

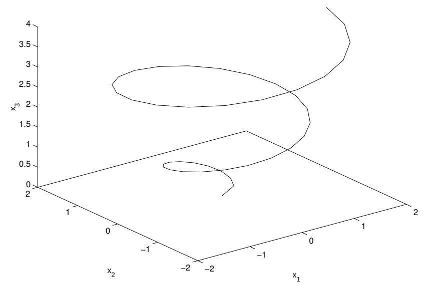


FIGURE 54.1. The curve $g : [0,4] \to \mathbb{R}^3$ with $g(t) = \big(t^{1/2}\cos(\pi t),\, t^{1/2}\sin(\pi t), t\big)$.

EXAMPLE 54.3. Let $f : [a,b] \to \mathbb{R}$ be given, and define $g : [a,b] \to \mathbb{R}^2$ by $g(t) = (g_1(t), g_2(t)) = (t, f(t))$. This curve is simply the graph of the function $f : [a,b] \to \mathbb{R}$, see Fig. 54.2.

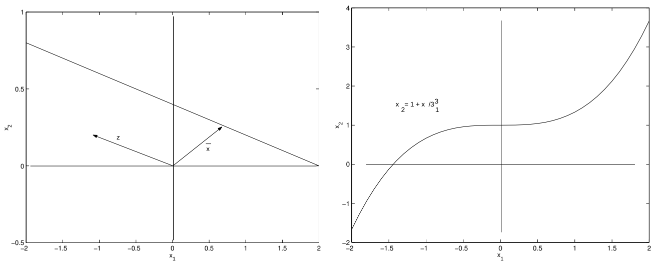


FIGURE 54.2. On the left: the curve $g(t) = \bar{x} + ta$. On the right: a curve $g(t) = (t, f(t))$.

54.3 Different parameterizations of a curve

It is possible to use different parametrizations for the set of points forming a curve. If $h : [c,d] \to [a,b]$ is a one-to-one mapping, then the composite function $f = g \circ h : [c,d] \to \mathbb{R}^2$ is a *reparameterization* of the curve $\{g(t) : t \in [a,b]\}$ given by $g : [a,b] \to \mathbb{R}^2$.

EXAMPLE 54.4. The function $f : [0,\infty) \to \mathbb{R}^3$ given by

$$f(\tau) = (\tau\cos(\pi\tau^2), \tau\sin(\pi\tau^2), \tau^2),$$

is a reparameterization of the curve $g : [0,\infty) \to \mathbb{R}^3$ given by

$$g(t) = (\sqrt{t}\cos(\pi t), \sqrt{t}\sin(\pi t), t),$$

obtained setting $t = h(\tau) = \tau^2$. We have $f = g \circ h$.

54.4 Surfaces in $\mathbb{R}^n$, $n \geq 3$

A function $g : Q \to \mathbb{R}^n$, where $n \geq 3$ and Q is a subdomain of $\mathbb{R}^2$, may be viewed to be a *surface* S in $\mathbb{R}^n$, see Fig. 54.3. We write $g = g(y)$ with $y = (y_1, y_2) \in Q$ and say that S is parameterized by $y \in Q$. We may also identify the surface S with the set of points $S = \{g(y) \in \mathbb{R}^n : y \in Q\}$, and reparameterize S by $f = g \circ h : \tilde{Q} \to \mathbb{R}^n$ if $h : \tilde{Q} \to Q$ is a one-to-one mapping of a domain $\tilde{Q}$ in $\mathbb{R}^2$ onto Q.

EXAMPLE 54.5. The simplest example of a surface $g : \mathbb{R}^2 \to \mathbb{R}^3$ is a plane in $\mathbb{R}^3$ given by

$$g(y) = g(y_1, y_2) = \bar{x} + y_1 b_1 + y_2 b_2, \quad y \in \mathbb{R}^2,$$

where $\bar{x}$, b_1, $b_2 \in \mathbb{R}^3$.

EXAMPLE 54.6. Let $f : [0,1] \times [0,1] \to \mathbb{R}$ be given, and define $g : [0,1] \times [0,1] \to \mathbb{R}^3$ by $g(y_1, y_2) = (y_1, y_2, f(y_1, y_2))$. This is a surface, which is the graph of $f : [0,1] \times [0,1] \to \mathbb{R}$. We also refer to this surface briefly as the surface given by the function $x_3 = f(x_1, x_2)$ with $(x_1, x_2) \in [0,1] \times [0,1]$.

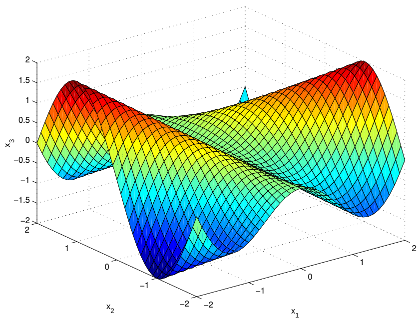


FIGURE 54.3. The surface $s(y_1, y_2) = \big(y_1, y_2, y_1 \sin\big((y_1 + y_2)\pi/2\big)\big)$ with $-1 \leq y_1, y_2 \leq 1$, or briefly the surface $x_3 = x_1 \sin\big((x_1 + x_2)\pi/2\big)$ with $-1 \leq x_1, x_2 \leq 1$.

54.5 Lipschitz continuity

We say that $f : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous on $\mathbb{R}^n$ if there is a constant L such that

$$\|f(x) - f(y)\| \leq L\|x - y\| \quad \text{for all } x, y \in \mathbb{R}^n. \tag{54.1}$$

This definition extends easily to functions $f : A \to \mathbb{R}^m$ with the domain $D(f) = A$ being a subset of $\mathbb{R}^n$. For example, A may be the unit n-cube $[0,1]^n = \{x \in \mathbb{R}^n : 0 \leq x_i \leq 1, i = 1, ..., n\}$ or the unit n-disc $\{x \in \mathbb{R}^n : \|x\| \leq 1\}$.

To check if a function $f : A \to \mathbb{R}^m$ is Lipschitz continuous on some subset A of $\mathbb{R}^n$, it suffices to check that the component functions $f_i : A \to \mathbb{R}$ are Lipschitz continuous. This is because

$$|f_i(x) - f_i(y)| \leq L_i\|x - y\| \quad \text{for } i = 1, ..., m,$$

implies

$$\|f(x) - f(y)\|^2 = \sum_{i=1}^{m} |f_i(x) - f_i(y)|^2 \leq \sum_{i=1}^{m} L_i^2 \|x - y\|^2,$$

which shows that $\|f(x) - f(y)\| \leq L\|x - y\|$ with $L = (\sum_i L_i^2)^{\frac{1}{2}}$.

Example 54.7. The function $f : [0,1] \times [0,1] \to \mathbb{R}^2$ defined by $f(x_1, x_2) = (x_1 + x_2, x_1 x_2)$, is Lipschitz continuous with Lipschitz constant $L = 2$. To show this, we note that $f_1(x_1, x_2) = x_1 + x_2$ is Lipschitz continuous on $[0,1] \times [0,1]$ with Lipschitz constant $L_1 = \sqrt{2}$ because $|f_1(x_1, x_2) - f_1(y_1, y_2)| \leq |x_1 - y_1| + |x_2 - y_2| \leq \sqrt{2}\|x - y\|$ by Cauchy's inequality. Similarly, $f_2(x_1, x_2) = x_1 x_2$ is Lipschitz continuous on $[0,1] \times [0,1]$ with Lipschitz constant $L_2 = \sqrt{2}$ since $|x_1 x_2 - y_1 y_2| = |x_1 x_2 - y_1 x_2 + y_1 x_2 - y_1 y_2| \leq |x_1 - y_1| + |x_2 - y_2| \leq \sqrt{2}\|x - y\|$.

Example 54.8. The function $f : \mathbb{R}^n \to \mathbb{R}^n$ defined by

$$f(x_1, ..., x_n) = (x_n, x_{n-1}, ..., x_1),$$

is Lipschitz continuous with Lipschitz constant $L = 1$.

Example 54.9. A linear transformation $f : \mathbb{R}^n \to \mathbb{R}^m$ given by an $m \times n$ matrix $A = (a_{ij})$, with $f(x) = Ax$ and x a n-column vector, is Lipschitz continuous with Lipschitz constant $L = \|A\|$. We made this observation in Chapter *Analytic geometry in* $\mathbb{R}^n$. We repeat the argument:

$$\begin{aligned} L = \max_{x \neq y} \frac{\|f(x) - f(y)\|}{\|x - y\|} &= \max_{x \neq y} \frac{\|Ax - Ay\|}{\|x - y\|} \\ &= \max_{x \neq y} \frac{\|A(x - y)\|}{\|x - y\|} = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \|A\|. \end{aligned}$$

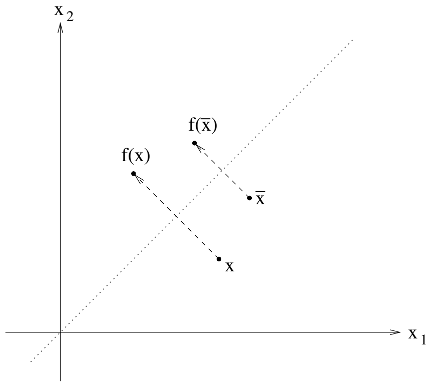


FIGURE 54.4. Illustration of the mapping $f(x_1, x_2) = (x_2, x_1)$, which is clearly Lipschitz continuous with $L = 1$.

Concerning the definition of the matrix norm $\|A\|$, we note that the function $F(x) = \|Ax\|/\|x\|$ is homogeneous of degree zero, that is, $F(\lambda x) = F(x)$ for all non-zero real numbers λ, and thus $\|A\|$ is the maximum value of $F(x)$ on the closed and bounded set $\{x \in \mathbb{R}^n : \|x\| = 1\}$, which is a finite real number.

We recall that if A is a diagonal $n \times n$ matrix with diagonal elements λ_i, then $\|A\| = \max_i |\lambda_i|$.

54.6 Differentiability: Jacobian, gradient and tangent

We say that $f : \mathbb{R}^n \to \mathbb{R}^m$ is *differentiable at* $\bar{x} \in \mathbb{R}^n$ if there is a $m \times n$ matrix $M(\bar{x}) = (m_{ij}(\bar{x}))$, called the *Jacobian* of the function $f(x)$ at $\bar{x}$, and a constant $K_f(\bar{x})$ such that for all x close to $\bar{x}$,

$$f(x) = f(\bar{x}) + M(\bar{x})(x - \bar{x}) + E_f(x, \bar{x}), \tag{54.2}$$

where $E_f(x, \bar{x}) = (E_f(x, \bar{x})_i)$ is an m-vector satisfying $\|E_f(x, \bar{x})\| \leq K_f(\bar{x})\|x - \bar{x}\|^2$. We also denote the Jacobian by $Df(\bar{x})$ or $f'(\bar{x})$ so that $M(\bar{x}) = Df(\bar{x}) = f'(\bar{x})$. Since $f(x)$ is a m-column vector, or $m \times 1$ matrix, and x is a n-column vector, or $n \times 1$ matrix, $M(\bar{x})(x - \bar{x})$ is the product of the $m \times n$ matrix $M(\bar{x})$ and the $n \times 1$ matrix $x - \bar{x}$ yielding a $m \times 1$ matrix or a m-column vector.

We say that $f : A \to \mathbb{R}^m$, where A is a subset of $\mathbb{R}^n$, is *differentiable on* A if $f(x)$ is differentiable at $\bar{x}$ for all $\bar{x} \in A$. We say that $f : A \to \mathbb{R}^m$ is *uniformly differentiable on* A if the constant $K_f(\bar{x}) = K_f$ can be chosen independently of $\bar{x} \in A$.

We now show how to determine a specific element $m_{ij}(\bar{x})$ of the Jacobian using the relation (54.2). We consider the coordinate function $f_i(x_1, ..., x_n)$ and setting $x = \bar{x} + se_j$, where e_j is the j^{th} standard basis vector and s is a small real number, we focus on the variation of $f_i(x_1, ..., x_n)$ as the variable x_j varies in a neighborhood of $\bar{x}_j$. The relation (54.2) states that for small non-zero real numbers s,

$$f_i(\bar{x} + se_j) = f_i(\bar{x}) + m_{ij}(\bar{x})s + E_f(\bar{x} + se_j, \bar{x})_i, \tag{54.3}$$

where $\|x - \bar{x}\|^2 = \|se_j\|^2 = s^2$ implies

$$|E_f(\bar{x} + se_j, \bar{x})_i| \leq K_f(\bar{x})s^2.$$

Note that by assumption $\|E_f(x, \bar{x})\| \leq K_f(\bar{x})\|x - \bar{x}\|^2$, and so each coordinate function $E_f(\bar{x} + se_j, \bar{x})_i$ satisfies $|E_f(x, \bar{x})_i| \leq K_f(\bar{x})\|x - \bar{x}\|^2$.

Now, dividing by s in (54.3) and letting s tend to zero, we find that

$$m_{ij}(\bar{x}) = \lim_{s \to 0} \frac{f_i(\bar{x} + se_j) - f_i(\bar{x})}{s}, \tag{54.4}$$

which we can also write as

$$\begin{aligned} &m_{ij}(\bar{x}) \\ &= \lim_{x_j \to \bar{x}_j} \frac{f_i(\bar{x}_1, ..., \bar{x}_{j-1}, x_j, \bar{x}_{j+1}, ..., \bar{x}_n) - f_i(\bar{x}_1, ..., \bar{x}_{j-1}, \bar{x}_j, \bar{x}_{j+1}, ..., \bar{x}_n)}{x_j - \bar{x}_j}. \end{aligned} \tag{54.5}$$

We refer to $m_{ij}(\bar{x})$ as the *partial derivative* of f_i with respect to x_j at $\bar{x}$, and we use the alternative notation $m_{ij}(\bar{x}) = \frac{\partial f_i}{\partial x_j}(\bar{x})$. To compute $\frac{\partial f_i}{\partial x_j}(\bar{x})$ we freeze all coordinates at $\bar{x}$ but the coordinate x_j and then let x_j vary in a neighborhood of $\bar{x}_j$. The formula

$$\begin{aligned} &\frac{\partial f_i}{\partial x_j}(\bar{x}) \\ &= \lim_{x_j \to \bar{x}_j} \frac{f_i(\bar{x}_1, ..., \bar{x}_{j-1}, x_j, \bar{x}_{j+1}, ..., \bar{x}_n) - f_i(\bar{x}_1, ..., \bar{x}_{j-1}, \bar{x}_j, \bar{x}_{j+1}, ..., \bar{x}_n)}{x_j - \bar{x}_j}, \end{aligned} \tag{54.6}$$

states that we compute the partial derivative with respect to the variable x_j by keeping all the other variables $x_1,...,x_{j-1}$, $x_{j+1},...,x_n$ constant. Thus, computing partial derivatives should be a pleasure using our previous expertise of computing derivatives of functions of one real variable!

We may express the computation alternatively as follows:

$$\frac{\partial f_i}{\partial x_j}(\bar{x}) = m_{ij}(\bar{x}) = g'_{ij}(0) = \frac{dg_{ij}}{ds}(0), \tag{54.7}$$

where $g_{ij}(s) = f_i(\bar{x} + se_j)$.

EXAMPLE 54.10. Let $f : \mathbb{R}^3 \to \mathbb{R}$ be given by $f(x_1, x_2, x_3) = x_1 e^{x_2} \sin(x_3)$. We compute

$$\frac{\partial f}{\partial x_1}(\bar{x}) = e^{\bar{x}_2} \sin(\bar{x}_3), \quad \frac{\partial f}{\partial x_2}(\bar{x}) = \bar{x}_1 e^{\bar{x}_2} \sin(\bar{x}_3),$$
$$\frac{\partial f}{\partial x_3}(\bar{x}) = \bar{x}_1 e^{\bar{x}_2} \cos(\bar{x}_3),$$

and thus

$$f'(\bar{x}) = (e^{\bar{x}_2} \sin(\bar{x}_3), \bar{x}_1 e^{\bar{x}_2} \sin(\bar{x}_3), \bar{x}_1 e^{\bar{x}_2} \cos(\bar{x}_3))$$

EXAMPLE 54.11. If $f : \mathbb{R}^3 \to \mathbb{R}^2$ is given by $f(x) = \begin{pmatrix} \exp(x_1^2 + x_2^2) \\ \sin(x_2 + 2x_3) \end{pmatrix}$, then

$$f'(x) = \begin{pmatrix} 2x_1 \exp(x_1^2 + x_2^2) & 2x_2 \exp(x_1^2 + x_2^2) & 0 \\ 0 & \cos(x_2 + 2x_3) & 2\cos(x_2 + 2x_3) \end{pmatrix}.$$

We have now shown how to compute the elements of a Jacobian using the usual rules for differentiation with respect to one real variable. This opens a whole new world of applications to explore. The setting is thus a differentiable function $f : \mathbb{R}^n \to \mathbb{R}^m$ satisfying for suitable $x, \bar{x} \in \mathbb{R}^n$:

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + E_f(x, \bar{x}), \tag{54.8}$$

with $\|E_f(x, \bar{x})\| \leq K_f(\bar{x})\|x - \bar{x}\|^2$, where $f'(\bar{x}) = Df(\bar{x})$ is the Jacobian $m \times n$ matrix with elements $\frac{\partial f_i}{\partial x_j}$:

$$f'(\bar{x}) = Df(\bar{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\bar{x}) & \frac{\partial f_1}{\partial x_2}(\bar{x}) & \dots & \frac{\partial f_1}{\partial x_n}(\bar{x}) \\ \frac{\partial f_2}{\partial x_1}(\bar{x}) & \frac{\partial f_2}{\partial x_2}(\bar{x}) & \dots & \frac{\partial f_2}{\partial x_n}(\bar{x}) \\ \dots & \dots & \dots & \\ \frac{\partial f_m}{\partial x_1}(\bar{x}) & \frac{\partial f_m}{\partial x_2}(\bar{x}) & \dots & \frac{\partial f_m}{\partial x_n}(\bar{x}) \end{pmatrix}.$$

Sometimes we use the following notation for the Jacobian $f'(x)$ of a function $y = f(x)$ with $f : \mathbb{R}^n \to \mathbb{R}^m$:

$$f'(x) = \frac{dy_1, \dots, dy_m}{dx_1, \dots, dx_n}(x) \tag{54.9}$$

The function $x \to \hat{f}(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x})$ is called the *linearization* of the function $x \to f(x)$ at $x = \bar{x}$. We have

$$\hat{f}(x) = f'(\bar{x})x + f(\bar{x}) - f'(\bar{x})\bar{x} = Ax + b,$$

with $A = f'(\bar{x})$ a $m \times n$ matrix and $b = f(\bar{x}) - f'(\bar{x})\bar{x}$ a m-column vector. We say that $\hat{f}(x)$ is an *affine transformation*, which is a transformation of the form $x \to Ax + b$, where x is a n-column vector, A is a $m \times n$ matrix and b is a m-column vector. The Jacobian $\hat{f}'(x)$ of the linearization $\hat{f}(x) = Ax + b$ is a constant matrix equal to the matrix A, because the partial derivatives of Ax with respect to x are simply the elements of the matrix A.

If $f : \mathbb{R}^n \to \mathbb{R}$, that is $m = 1$, then we also denote the Jacobian f' by ∇f, that is,

$$f'(\bar{x}) = \nabla f(\bar{x}) = (\frac{\partial f}{\partial x_1}(\bar{x}), ..., \frac{\partial f}{\partial x_n}(\bar{x})).$$

In words, $\nabla f(\bar{x})$ is the n-row vector or $1 \times n$ matrix of partial derivatives of $f(x)$ with respect to x_1, x_2,...,x_n at $\bar{x}$. We refer to $\nabla f(\bar{x})$ as the *gradient* of $f(x)$ at $\bar{x}$. If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\bar{x}$, we thus have

$$f(x) = f(\bar{x}) + \nabla f(\bar{x})(x - \bar{x}) + E_f(x, \bar{x}), \qquad (54.10)$$

with $|E_f(x, \bar{x})| \leq K_f(\bar{x})\|x - \bar{x}\|^2$, and $\hat{f}(x) = f(\bar{x}) + \nabla f(\bar{x})(x - \bar{x})$ is the linearization of $f(x)$ at $x = \bar{x}$. We may alternatively express the product $\nabla f(\bar{x})(x - \bar{x})$ of the n-row vector ($1 \times n$ matrix) $\nabla f(\bar{x})$ with the n-column vector ($n \times 1$ matrix) $(x - \bar{x})$ as the scalar product $\nabla f(\bar{x}) \cdot (x - \bar{x})$ of the n-vector $\nabla f(\bar{x})$ with the n-vector $(x - \bar{x})$. We thus often write (54.10) in the form

$$f(x) = f(\bar{x}) + \nabla f(\bar{x}) \cdot (x - \bar{x}) + E_f(x, \bar{x}). \qquad (54.11)$$

Example 54.12. If $f : \mathbb{R}^3 \to \mathbb{R}$ is given by $f(x) = x_1^2 + 2x_2^3 + 3x_3^4$, then

$$\nabla f(x) = (2x_1, 6x_2^2, 12x_3^3).$$

Example 54.13. The equation $x_3 = f(x)$ with $f : \mathbb{R}^2 \to \mathbb{R}$ and $x = (x_1, x_2)$ represents a surface in $\mathbb{R}^3$ (the graph of the function f). The linearization

$$x_3 = f(\bar{x}) + \nabla f(\bar{x}) \cdot (x - \bar{x}) = f(\bar{x}) + \frac{\partial f}{\partial x_1}(\bar{x})(x_1 - \bar{x}_1) + \frac{\partial f}{\partial x_2}(\bar{x})(x_2 - \bar{x}_2)$$

with $\bar{x} = (\bar{x}_1, \bar{x}_2)$, represents the *tangent plane* at $x = \bar{x}$, see Fig. 54.13.

Example 54.14. Consider now a curve $f : \mathbb{R} \to \mathbb{R}^m$, that is, $f(t) = (f_1(t), ..., f_m(t))$ with $t \in \mathbb{R}$ and we have a situation with $n = 1$. The linearization $t \to \hat{f}(t) = f(\bar{t}) + f'(\bar{t})(t - \bar{t})$ at $\bar{t}$ represents a straight line in $\mathbb{R}^m$ through the point $f(\bar{t})$ and the Jacobian $f'(\bar{t}) = (f_1'(\bar{t}), ..., f_m'(\bar{t}))$ gives the direction of the *tangent* to the curve $f : \mathbb{R} \to \mathbb{R}^m$ at $f(\bar{t})$, see Fig. 54.6.

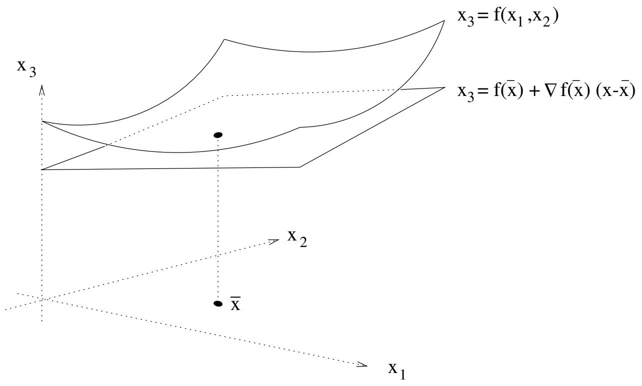


FIGURE 54.5. The surface $x_3 = f(x_1, x_2)$ and its tangent plane.

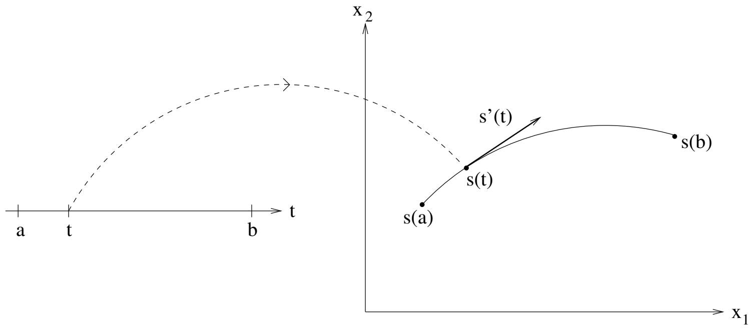


FIGURE 54.6. The tangent $s'(t)$ to a curve given by $s(t)$.

54.7 The Chain rule

Let $g : \mathbb{R}^n \to \mathbb{R}^m$ and $f : \mathbb{R}^m \to \mathbb{R}^p$ and consider the composite function $f \circ g : \mathbb{R}^n \to \mathbb{R}^p$ defined by $f \circ g(x) = f(g(x))$. Under suitable assumptions of differentiability and Lipschitz continuity, we shall prove a *Chain rule* generalizing the Chain rule of Chapter *Differentiation rules* in the case $n = m = p = 1$. Using linearizations of f and g, we have

$$\begin{aligned} f(g(x)) &= f(g(\bar{x})) + f'(g(\bar{x}))(g(x) - g(\bar{x})) + E_f(g(x), g(\bar{x})) \\ &= f(g(\bar{x})) + f'(g(\bar{x}))g'(\bar{x})(x - \bar{x}) + f'(g(\bar{x}))E_g(x, \bar{x}) + E_f(g(x), g(\bar{x})), \end{aligned}$$

where we may naturally assume that $\|E_f(g(x), g(\bar{x}))\| \leq K_f \|g(x) - g(\bar{x})\|^2 \leq K_f L_g^2 \|x - \bar{x}\|^2$, and $\|f'(g(\bar{x}))E_g(x, \bar{x})\| \leq \|f'(g(\bar{x}))\| K_g \|x - \bar{x}\|^2$, with suitable constants of differentiability K_f and K_g and Lipschitz constant L_g. We have now proved

Theorem 54.1 (The Chain rule) *If $g : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\bar{x} \in \mathbb{R}^n$, and $f : \mathbb{R}^m \to \mathbb{R}^p$ is differentiable at $g(\bar{x}) \in \mathbb{R}^m$ and further $g : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous, then the composite function $f \circ g : \mathbb{R}^n \to \mathbb{R}^p$ is differentiable at $\bar{x} \in \mathbb{R}^n$ with Jacobian*

$$(f \circ g)'(\bar{x}) = f'(g(\bar{x}))g'(\bar{x}).$$

The Chain rule has a wealth of applications and we now turn to harvest a couple of the most basic examples.

54.8 The Mean Value theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable on $\mathbb{R}^n$ with a Lipschitz continuous gradient, and for given $x, \bar{x} \in \mathbb{R}^n$ consider the function $h : \mathbb{R} \to \mathbb{R}$ defined by

$$h(t) = f(\bar{x} + t(x - \bar{x})) = f \circ g(t),$$

with $g(t) = \bar{x} + t(x - \bar{x})$ representing the straight line through $\bar{x}$ and x. We have

$$f(x) - f(\bar{x}) = h(1) - h(0) = h'(\bar{t}),$$

for some $\bar{t} \in [0, 1]$, where we applied the usual Mean Value theorem to the function $h(t)$. By the Chain rule we have

$$h'(t) = \nabla f(g(t)) \cdot g'(t) = \nabla f(g(t)) \cdot (x - \bar{x}),$$

and we have now proved:

Theorem 54.2 (Mean Value theorem) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable on $\mathbb{R}^n$ with a Lipschitz continuous gradient ∇f. Then for given x and $\bar{x}$ in $\mathbb{R}^n$, there is $y = x + \bar{t}(x - \bar{x})$ with $\bar{t} \in [0, 1]$, such that*

$$f(x) - f(\bar{x}) = \nabla f(y) \cdot (x - \bar{x}).$$

With the help of the Mean Value theorem we express the difference $f(x) - f(\bar{x})$ as the scalar product of the gradient $\nabla f(y)$ with the difference $x - \bar{x}$, where y is a point somewhere on the straight line between x and $\bar{x}$.

We may extend the Mean Value theorem to a function $f : \mathbb{R}^n \to \mathbb{R}^m$ to take the form

$$f(x) - f(\bar{x}) = f'(y)(x - \bar{x}),$$

where y is a point on the straight line between x and $\bar{x}$, which may be different for different rows of $f'(y)$. We may then estimate:

$$\|f(x) - f(\bar{x})\| = \|f'(y) \cdot (x - \bar{x})\| \leq \|f'(y)\| \|x - \bar{x}\|,$$

and we may thus estimate the Lipschitz constant of f by $\max_y \|f'(y)\|$ with $\|f'(y)\|$ the (Euclidean) matrix norm of $f'(y)$.

EXAMPLE 54.15. Let $f : \mathbb{R}^n \to \mathbb{R}$ be given by $f(x) = \sin(\sum_{j=1}^n x_j)$. We have

$$\frac{\partial f}{\partial x_i}(\bar{x}) = \cos(\sum_{j=1}^n \bar{x}_j) \quad \text{for } i = 1, ..., n,$$

and thus $|\frac{\partial f}{\partial x_i}(\bar{x})| \leq 1$ for $i = 1, ..., n$, and therefore

$$\|\nabla f(\bar{x})\| \leq \sqrt{n}.$$

We conclude that $f(x) = \sin(\sum_{j=1}^n x_j)$ is Lipschitz continuous with Lipschitz constant $\sqrt{n}$.

54.9 Direction of steepest descent and the gradient

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a given function and suppose we want to study the variation of $f(x)$ in a neighborhood of a given point $\bar{x} \in \mathbb{R}^n$. More precisely, let x vary on the line through $\bar{x}$ in a given direction $z \in \mathbb{R}^n$, that is assume that $x = \bar{x} + tz$ where t is a real variable varying in a neighborhood of 0. Assuming f to be differentiable, the linearization formula (54.8) implies

$$f(x) = f(\bar{x}) + t\nabla f(\bar{x}) \cdot z + E_f(x, \bar{x}), \tag{54.12}$$

where $|E_f(x, \bar{x})| \leq t^2 K_f \|z\|^2$ and $\nabla f(\bar{x}) \cdot z$ is the scalar product of the gradient $\nabla f(\bar{x}) \in \mathbb{R}^n$ and the vector $z \in \mathbb{R}^n$. If $\nabla f(\bar{x}) \cdot z \neq 0$, then the linear term $t\nabla f(\bar{x}) \cdot z$ will dominate the quadratic term $E_f(x, \bar{x})$ for small t. So the linearization

$$\hat{f}(x) = f(\bar{x}) + t\nabla f(\bar{x}) \cdot z$$

will be a good approximation of $f(x)$ for $x = \bar{x} + tz$ close to $\bar{x}$. Thus if $\nabla f(\bar{x}) \cdot z \neq 0$, then we get good information on the variation of $f(x)$ along the line $x = \bar{x} + tz$ by studying the linear function $t \to f(\bar{x}) + t\nabla f(\bar{x}) \cdot z$ with slope $\nabla f(\bar{x}) \cdot z$. In particular, if $\nabla f(\bar{x}) \cdot z > 0$ and $x = \bar{x} + tz$ then $\hat{f}(x)$ increases as we increase t and decreases as we decrease t. Similarly, if $\nabla f(\bar{x}) \cdot z < 0$ and $x = \bar{x} + tz$ then $\hat{f}(x)$ decreases as we increase t and increases as we decrease t.

Alternatively, we may consider the composite function $F_z : \mathbb{R} \to \mathbb{R}$ defined by $F_z(t) = f(g_z(t))$ with $g_z : \mathbb{R} \to \mathbb{R}^n$ given by $g_z(t) = \bar{x} + tz$. Obviously, $F_z(t)$ describes the variation of $f(x)$ on the straight line through $\bar{x}$ with direction z, with $F_z(0) = f(\bar{x})$. Of course, the derivative $F_z'(0)$ gives important information on this variation close to $\bar{x}$. By the Chain rule we have

$$F_z'(0) = \nabla f(\bar{x})z = \nabla f(\bar{x}) \cdot z,$$

and we retrieve $\nabla f(\bar{x}) \cdot z$ as a quantity of interest. In particular, the sign of $\nabla f(\bar{x}) \cdot z$ determines if $F_z(t)$ is increasing or decreasing at $t = 0$.

We may now ask how to choose the direction z to get maximal increase or decrease. We assume $\nabla f(\bar{x}) \neq 0$ to avoid the trivial case with $F_z'(0) = 0$ for all z. It is then natural to normalize z so $\|z\| = 1$ and we study the quantity $F_z'(0) = \nabla f(\bar{x}) \cdot z$ as we vary $z \in \mathbb{R}^n$ with $\|z\| = 1$. We conclude that the scalar product $\nabla f(\bar{x}) \cdot z$ is maximized if we choose z in the direction of the gradient $\nabla f(\bar{x})$,

$$z = \frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|},$$

which is called the direction of *steepest ascent*. For this gives

$$\max_{\|z\|=1} F_z'(0) = \nabla f(\bar{x}) \cdot \frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|} = \|\nabla f(\bar{x})\|.$$

Similarly, the scalar product is minimized if we choose z in the opposite direction of the gradient $\nabla f(\bar{x})$,

$$z = -\frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|},$$

which is called the direction of *steepest descent*, see Fig. 54.7. For then

$$\min_{\|z\|=1} F_z'(0) = -\nabla f(\bar{x}) \cdot \frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|} = -\|\nabla f(\bar{x})\|.$$

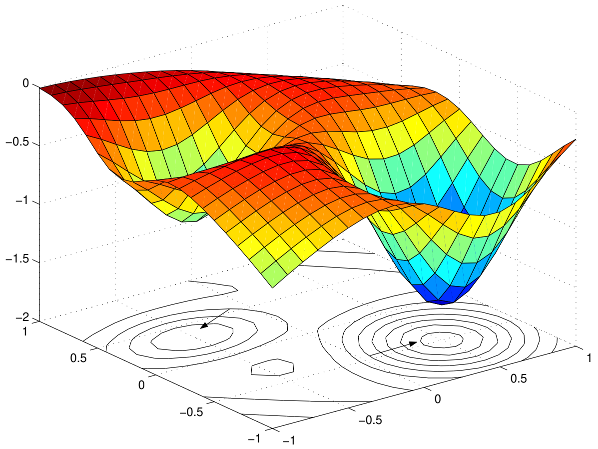


FIGURE 54.7. Directions of steepest descent on a "hiking map".

If $\nabla f(\bar{x}) = 0$, then $\bar{x}$ is said to be a *stationary point*. If $\bar{x}$ is a stationary point, then evidently $\nabla f(\bar{x}) \cdot z = 0$ for any direction z and

$$f(x) = f(\bar{x}) + E_f(x, \bar{x}).$$

The difference $f(x)-f(\bar{x})$ is then quadratically small in the distance $\|x-\bar{x}\|$, that is $|f(x) - f(\bar{x})| \leq K_f\|x - \bar{x}\|^2$, and $f(x)$ is very close to the constant value $f(\bar{x})$ for x close to $\bar{x}$.

54.10 A minimum point is a stationary point

Suppose $\bar{x} \in \mathbb{R}^n$ is a *minimum point* for the function $f : \mathbb{R}^n \to \mathbb{R}$, that is

$$f(x) \geq f(\bar{x}) \quad \text{for } x \in \mathbb{R}^n. \tag{54.13}$$

We shall show that if $f(x)$ is differentiable at a minimum point $\bar{x}$, then

$$\nabla f(\bar{x}) = 0. \tag{54.14}$$

For if $\nabla f(\bar{x}) \neq 0$, then we could move in the direction of steepest descent from $\bar{x}$ to a point x close to $\bar{x}$ with $f(x) < f(\bar{x})$, contradicting (54.13). Consequently, in order to find minimum points of a function $f : \mathbb{R}^n \to \mathbb{R}$, we are led to try to solve the equation $g(x) = 0$, where $g = \nabla f : \mathbb{R}^n \to \mathbb{R}^n$. Here, we interpret $\nabla f(x)$ as a n-column vector.

A whole world of applications in mechanics, physics and other areas may be formulated as solving equations of the form $\nabla f(x) = 0$, that is as finding stationary points. We shall meet many applications below.

54.11 The method of Steepest Descent to find a minimum point

Let $f : \mathbb{R}^n \to \mathbb{R}$ be given and consider the problem of finding a minimum point $\bar{x}$. To do so it is natural to try a *method of Steepest Descent*: Given an approximation $\bar{y}$ of $\bar{x}$ with $\nabla f(\bar{y}) \neq 0$, we move from $\bar{y}$ to a new point y in the direction of steepest descent:

$$y = \bar{y} - \alpha\frac{\nabla f(\bar{y})}{\|\nabla f(\bar{y})\|},$$

where $\alpha > 0$ is a step length to be chosen. We know that $f(y)$ decreases as α increases from 0 and the question is just to find a reasonable value of α. This can be done by increasing α in small steps until $f(y)$ doesn't decrease anymore. The procedure is then repeated with $\bar{y}$ replaced by y. Evidently, the method of Steepest Descent is closely connected to Fixed Point Iteration for solving the equation $\nabla f(x) = 0$ in the form

$$x = x - \alpha\nabla f(x)$$

where we let $\alpha > 0$ include the normalizing factor $1/\|\nabla f(\bar{y})\|$.

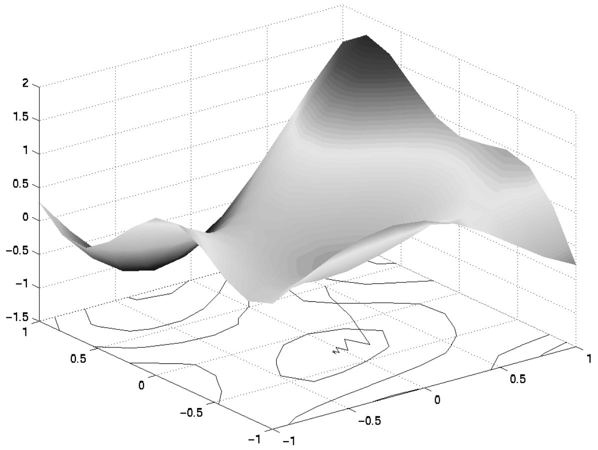


FIGURE 54.8. The method of Steepest Descent for $f(x_1, x_2) = x_1 \sin(x_1 + x_2) + x_2 \cos(2x_1 - 3x_2)$ starting at $(.5, .5)$ with $\alpha = .3$.

54.12 Directional derivative

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$, let $g_z(t) = \bar{x} + tz$ with $z \in \mathbb{R}^n$ a given vector normalized to $\|z\| = 1$, and consider the composite function $F_z(t) = f(\bar{x} + tz)$. The Chain rule implies

$$F_z'(0) = \nabla f(\bar{x}) \cdot z,$$

and

$$\nabla f(\bar{x}) \cdot z$$

is called the *derivative of* $f(x)$ *in the direction* z at $\bar{x}$, see Fig. 54.9.

54.13 Higher order partial derivatives

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable on $\mathbb{R}^n$. Each partial derivative $\frac{\partial f}{\partial x_i}(\bar{x})$ is a function of $\bar{x} \in \mathbb{R}^n$ may be itself be differentiable. We denote its partial derivatives by

$$\frac{\partial}{\partial x_j}\frac{\partial f}{\partial x_i}(\bar{x}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\bar{x}), \quad i, j = 1, ..., n, \, \bar{x} \in \mathbb{R}^n,$$

which are called the *partial derivatives of* f *of second order at* $\bar{x}$. It turns out that under appropriate continuity assumptions, the order of differentiation does not matter. In other words, we shall prove that

$$\frac{\partial^2 f}{\partial x_j \partial x_i}(\bar{x}) = \frac{\partial^2 f}{\partial x_i \partial x_j}(\bar{x}).$$

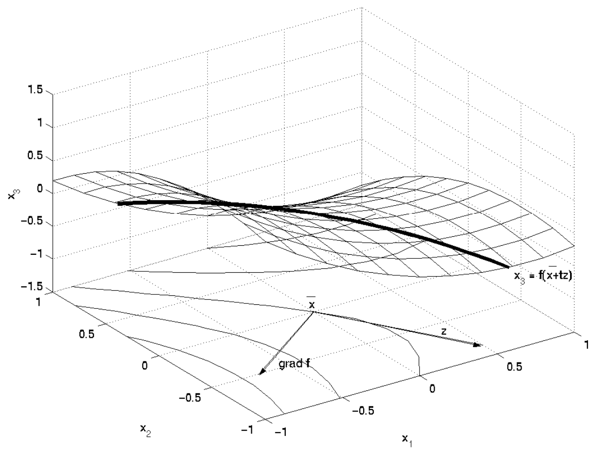


FIGURE 54.9. Illustration of directional derivative.

We carry out the proof in the case $n = 2$ with $i = 1$ and $j = 2$. We rewrite the expression

$$A = f(x_1, x_2) - f(\bar{x}_1, x_2) - f(x_1, \bar{x}_2) + f(\bar{x}_1, \bar{x}_2), \tag{54.15}$$

as

$$A = f(x_1, x_2) - f(x_1, \bar{x}_2) - f(\bar{x}_1, x_2) + f(\bar{x}_1, \bar{x}_2), \tag{54.16}$$

by shifting the order of the two mid terms. First, we set $F(x_1, x_2) = f(x_1, x_2) - f(\bar{x}_1, x_2)$ and use (54.15) to write

$$A = F(x_1, x_2) - F(x_1, \bar{x}_2).$$

The Mean Value theorem implies

$$A = \frac{\partial F}{\partial x_2}(x_1, y_2)(x_2 - \bar{x}_2) = \big(\frac{\partial f}{\partial x_2}(x_1, y_2) - \frac{\partial f}{\partial x_2}(\bar{x}_1, y_2)\big)(x_2 - \bar{x}_2)$$

for some $y_2 \in [\bar{x}_2, x_2]$. We use the Mean value theorem once again to get

$$A = \frac{\partial^2 f}{\partial x_1 \partial x_2}(y_1, y_2)(x_1 - \bar{x}_1)(x_2 - \bar{x}_2),$$

with $y_1 \in [\bar{x}_1, x_1]$. We next rewrite A using (54.16) in the form

$$A = G(x_1, x_2) - G(\bar{x}_1, x_2),$$

where $G(x_1, x_2) = f(x_1, x_2) - f(x_1, \bar{x}_2)$. Using the Mean Value theorem twice as above, we obtain

$$A = \frac{\partial^2 f}{\partial x_2 \partial x_1}(z_1, z_2)(x_1 - \bar{x}_1)(x_2 - \bar{x}_2),$$

where $z_i \in [\bar{x}_i, x_i]$, $i = 1, 2$. Assuming the second partial derivatives are Lipschitz continuous at $\bar{x}$ and letting x_i tend to $\bar{x}_i$ for $i = 1, 2$ gives

$$\frac{\partial^2 f}{\partial x_1 \partial x_2}(\bar{x}) = \frac{\partial^2 f}{\partial x_2 \partial x_1}(\bar{x}).$$

We have proved the following fundamental result.

Theorem 54.3 *If the partial derivatives of second order of a function $f : \mathbb{R}^n \to \mathbb{R}$ are all Lipschitz continuous, then the order of application of the derivatives of second order is irrelevant.*

The result directly generalizes to higher order partial derivatives: if the derivatives are Lipschitz continuous, then the order of application doesn't matter. What a relief!

54.14 Taylor's theorem

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ has Lipschitz continuous partial derivatives of order 2. For given $x, \bar{x} \in \mathbb{R}^n$, consider the function $h : \mathbb{R} \to \mathbb{R}$ defined by

$$h(t) = f(\bar{x} + t(x - \bar{x})) = f \circ g(t),$$

where $g(t) = \bar{x} + t(x - \bar{x})$ is the straight line through $\bar{x}$ and x. Clearly $h(1) = f(x)$ and $h(0) = f(\bar{x})$, so the Taylor's theorem applied to $h(t)$ gives

$$h(1) = h(0) + h'(0) + \frac{1}{2}h''(\bar{t}),$$

for some $\bar{t} \in [0, 1]$. We compute using the Chain rule:

$$h'(t) = \nabla f(g(t)) \cdot (x - \bar{x}) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(g(t))(x_i - \bar{x}_i),$$

and similarly by a further differentiation with respect to t:

$$h''(t) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(g(t))(x_i - \bar{x}_i)(x_j - \bar{x}_j).$$

We thus obtain

$$f(x) = f(\bar{x}) + \nabla f(\bar{x}) \cdot (x - \bar{x}) + \frac{1}{2} \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(y)(x_i - \bar{x}_i)(x_j - \bar{x}_j), \quad (54.17)$$

for some $y = \bar{x} + \bar{t}(x - \bar{x})$ with $t \in [0, 1]$. The $n \times n$ matrix $H(\bar{x}) = (h_{ij}(\bar{x}))$ with elements $h_{ij}(\bar{x}) = \frac{\partial^2 f}{\partial x_i \partial x_j}(\bar{x})$ is called the *Hessian* of $f(x)$ at $x = \bar{x}$.

The Hessian is the matrix of all second partial derivatives of $f : \mathbb{R}^n \to \mathbb{R}$. With matrix vector notation with x a n-column vector, we can write

$$\sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(y)(x_i - \bar{x}_i)(x_j - \bar{x}_j) = (x - \bar{x})^\top H(y)(x - \bar{x}).$$

We summarize:

Theorem 54.4 (Taylor's theorem) *Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable with Lipschitz continuous Hessian $H = (h_{ij})$ with elements $h_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$. Then, for given x and $\bar{x} \in \mathbb{R}^n$, there is $y = x + \bar{t}(x - \bar{x})$ with $\bar{t} \in [0, 1]$, such that*

$$\begin{aligned} f(x) &= f(\bar{x}) + \nabla f(\bar{x}) \cdot (x - \bar{x}) + \frac{1}{2} \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(y)(x_i - \bar{x}_i)(x_j - \bar{x}_j) \\ &= f(\bar{x}) + \nabla f(\bar{x}) \cdot (x - \bar{x}) + \frac{1}{2}(x - \bar{x})^\top H(y)(x - \bar{x}). \end{aligned}$$

54.15 Contraction mapping theorem

We shall now prove the following generalization of the Contraction Mapping theorem.

Theorem 54.5 *If $g : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with Lipschitz constant $L < 1$, then the equation $x = g(x)$ has a unique solution $\bar{x} = \lim_{i\to\infty} x^{(i)}$, where $\{x^{(i)}\}_{i=1}^{\infty}$ is a sequence in $\mathbb{R}^n$ generated by Fixed Point Iteration: $x^{(i)} = g(x^{(i-1)})$, $i = 1, 2, ...,$ starting with any initial value $x^{(0)}$.*

The proof is word by word the same as in the case $g : \mathbb{R} \to \mathbb{R}$ considered above. We repeat the proof for the convenience of the reader. Subtracting the equation $x^{(k)} = g(x^{(k-1)})$ from $x^{(k+1)} = g(x^{(k)})$, we get

$$x^{(k+1)} - x^{(k)} = g(x^{(k)}) - g(x^{(k-1)}),$$

and using the Lipschitz continuity of g, we thus have

$$\|x^{(k+1)} - x^{(k)}\| \leq L\|x^{(k)} - x^{(k-1)}\|. \tag{54.18}$$

Repeating this estimate, we find that

$$\|x^{(k+1)} - x^{(k)}\| \leq L^k\|x^{(1)} - x^{(0)}\|, \tag{54.19}$$

and thus for $j > i$

$$\begin{aligned} \|x^{(i)} - x^{(j)}\| &\leq \sum_{k=i}^{j-1} \|x^{(k)} - x^{(k+1)}\| \\ &\leq \|x^{(1)} - x^{(0)}\| \sum_{k=i}^{j-1} L^k = \|x^{(1)} - x^{(0)}\| L^i \, \frac{1 - L^{j-i}}{1 - L}. \end{aligned}$$

Since $L < 1$, $\{x^{(i)}\}_{i=1}^{\infty}$ is a Cauchy sequence in $\mathbb{R}^n$, and therefore converges to a limit $\bar{x} = \lim_{i\to\infty} x^{(i)}$. Passing to the limit in the equation $x^{(i)} = g(x^{(i-1)})$ shows that $\bar{x} = g(\bar{x})$ and thus $\bar{x}$ is a fixed point of $g : \mathbb{R}^n \to \mathbb{R}^n$. Uniqueness follows from the fact that if $\bar{y} = g(\bar{y})$, then $\|\bar{x} - \bar{y}\| = \|g(\bar{x}) - g(\bar{y})\| \le L\|\bar{x} - \bar{y}\|$ which is impossible unless $\bar{y} = \bar{x}$, because $L < 1$.

EXAMPLE 54.16. Consider the function $g : \mathbb{R}^2 \to \mathbb{R}^2$ defined by $g(x) = (g_1(x), g_2(x))$ with

$$g_1(x) = \frac{1}{4 + |x_1| + |x_2|}, \quad g_2(x) = \frac{1}{4 + |\sin(x_1)| + |\cos(x_2)|}.$$

We have

$$\left|\frac{\partial g_i}{\partial x_j}\right| \le \frac{1}{16},$$

and thus by simple estimates

$$\|g(x) - g(y)\| \le \frac{1}{4}\|x - y\|,$$

which shows that $g : \mathbb{R}^2 \to \mathbb{R}^2$ is Lipschitz continuous with Lipschitz constant $L_g \le \frac{1}{4}$. The equation $x = g(x)$ thus has a unique solution.

54.16 Solving $f(x) = 0$ with $f : \mathbb{R}^n \to \mathbb{R}^n$

The Contraction Mapping theorem can be applied as follows. Suppose $f : \mathbb{R}^n \to \mathbb{R}^n$ is given and we want to solve the equation $f(x) = 0$. Introduce

$$g(x) = x - Af(x),$$

where A is some non-singular $n \times n$ matrix with constant coefficients to be chosen. The equation $x = g(x)$ is then equivalent to the equation $f(x) = 0$. If $g : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with Lipschitz constant $L < 1$, then $g(x)$ has a unique fixed point $\bar{x}$ and thus $f(\bar{x}) = 0$. We have

$$g'(x) = I - Af'(x),$$

and thus we are led to choose the matrix A so that

$$\|I - Af'(x)\| \le 1$$

for x close to the root $\bar{x}$. The ideal choice seems to be:

$$A = f'(\bar{x})^{-1},$$

assuming that $f'(\bar{x})$ is non-singular, since then $g'(\bar{x}) = 0$. In applications, we may seek to choose A close to $f'(\bar{x})^{-1}$ with the hope that the corresponding $g'(x) = I - Af'(x)$ will have $\|g'x)\|$ small for x close to the root $\bar{x}$, leading to a quick convergence. In Newton's method we choose $A = f'(x)^{-1}$, see below.

EXAMPLE 54.17. Consider the initial value problem $\dot{u}(t) = f(u(t))$ for $t > 0$, $u(0) = u_0$, where $f : \mathbb{R}^n \to \mathbb{R}^n$ is a given Lipschitz continuous function with Lipschitz constant L_f, and as usual $\dot{u} = \frac{du}{dt}$. Consider the backward Euler method

$$U(t_i) = U(t_{i-1}) + k_i f(U(t_i)), \tag{54.20}$$

where $0 = t_0 < t_1 < t_2....$ is a sequence of increasing discrete time levels with time steps $k_i = t_i - t_{i-1}$. To determine $U(t_i) \in \mathbb{R}^n$ satisfying (54.20) having already determined $U(t_{i-1})$, we have to solve the nonlinear system of equations

$$V = U(t_{i-1}) + k_i f(V) \tag{54.21}$$

in the unknown $V \in \mathbb{R}^n$. This equation is of the form $V = g(V)$ with $g(V) = U(t_{i-1}) + k_i f(V)$ and $g : \mathbb{R}^n \to \mathbb{R}^n$.

Therefore, we use the Fixed Point Iteration

$$V^{(m)} = U(t_{i-1}) + k_i f(V^{(m-1)}), \quad m = 1, 2, ...,$$

choosing say $V^{(0)} = U(t_{i-1})$ to try to solve for the new value. If L_f denotes the Lipschitz constant of $f : \mathbb{R}^n \to \mathbb{R}^n$, then

$$\|g(V)-g(W)\| = \|k_i(f(V)-f(W))\| \leq k_i L_f \|V-W\|, \quad \text{for all } V, W \in \mathbb{R}^n,$$

and thus $g : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with Lipschitz constant $L_g = k_i L_f$. Now $L_g < 1$ if the time step k_i satisfies $k_i < 1/L_f$ and thus the Fixed Point Iteration to determine $U(t_i)$ in (54.20) converges if $k_i < 1/L_f$. This gives a method for numerical solution of a very large class of initial value problems of the form $\dot{u}(t) = f(u(t))$ for $t > 0$, $u(0) = u_0$. The only restriction is to choose sufficiently small time steps, can be a severe restriction if the Lipschitz constant L_f is very large in the sense of requiring massive computational work (very small time steps). Thus, caution for large Lipschitz constants L_f!!

54.17 The Inverse Function theorem

Suppose $f : \mathbb{R}^n \to \mathbb{R}^n$ is a given function and let $\bar{y} = f(\bar{x})$, where $\bar{x} \in \mathbb{R}^n$ is given. We shall prove that if $f'(\bar{x})$ is non-singular, then for $y \in \mathbb{R}^n$ close to $\bar{y}$, the equation

$$f(x) = y \tag{54.22}$$

has a unique solution x. Thus, we can define x as a function of y for y close to $\bar{y}$, which is called the inverse function $x = f^{-1}(y)$ of $y = f(x)$. To show that (54.22) has a unique solution x for any given y close to $\bar{y}$, we consider the Fixed Point iteration for $x = g(x)$ with $g(x) = x - (f'(\bar{x}))^{-1}(f(x) - y)$, which has the fixed point x satisfying $f(x) = y$ as desired. The iteration is

$$x^{(j)} = x^{(j-1)} - (f'(\bar{x}))^{-1}(f(x^{(j-1)}) - y), \quad j = 1, 2, \ldots,$$

with $x^{(0)} = \bar{x}$. To analyze the convergence, we subtract

$$x^{(j-1)} = x^{(j-2)} - (f'(\bar{x}))^{-1}(f(x^{(j-2)}) - y)$$

and write $e^j = x^{(j)} - x^{(j-1)}$ to get

$$e^j = e^{j-1} - (f'(\bar{x}))^{-1}(f(x^{(j-1)}) - f(x^{(j-2)}) \quad \text{for } j = 1, 2, \ldots$$

The Mean Value theorem implies

$$f_i(x^{(j-1)}) - f_i(x^{(j-2)}) = f'(z)e^{j-1},$$

where z lies on the straight line between $x^{(j-1)}$ and $x^{(j-2)}$. Note there might be possibly different z for different rows of $f'(z)$. We conclude that

$$e^j = \left(I - (f'(\bar{x}))^{-1} f'(z)\right) e^{j-1}.$$

Assuming now that

$$\|I - (f'(\bar{x}))^{-1} f'(z)\| \leq \theta, \tag{54.23}$$

where $\theta < 1$ is a positive constant, we have

$$\|e^j\| \leq \theta \|e^{j-1}\|.$$

As in the proof of the Contraction Mapping theorem, this shows that the sequence $\{x^{(j)}\}_{j=1}^{\infty}$ is a Cauchy sequence and thus converges to a vector $x \in \mathbb{R}^n$ satisfying $f(x) = y$.

The condition for convergence is obviously (54.23). This condition is satisfied if the coefficients of the Jacobian $f'(x)$ are Lipschitz continuous close to $\bar{x}$ and $f'(\bar{x})$ is non-singular so that $(f'(\bar{x}))^{-1}$ exists, and we restrict y to be sufficiently close to $\bar{y}$.

We summarize in the following (very famous):

Theorem 54.6 (Inverse Function theorem) *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ and assume the coefficients of $f'(x)$ are Lipschitz continuous close to $\bar{x}$ and $f'(\bar{x})$ is non-singular. Then for y sufficiently close to $\bar{y} = f(\bar{x})$, the equation $f(x) = y$ has a unique solution x. This defines x as a function $x = f^{-1}(y)$ of y.*

54.18 The Implicit Function theorem

There is an important generalization of the Inverse Function theorem. Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ be a given function with value $f(x, y) \in \mathbb{R}^n$ for $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. Assume that $f(\bar{x}, \bar{y}) = 0$ and consider the equation in $x \in \mathbb{R}^n$,

$$f(x, y) = 0,$$

for $y \in \mathbb{R}^m$ close to $\bar{y}$. In the case of the Inverse Function theorem we considered a special case of this situation with $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ defined by $f(x,y) = g(x) - y$ with $g : \mathbb{R}^n \to \mathbb{R}^n$.

We define the Jacobian $f'_x(x,y)$ of $f(x,y)$ with respect to x at (x,y) to be the $n \times n$ matrix with elements

$$\frac{\partial f_i}{\partial x_j}(x,y).$$

Assuming now that $f'_x(\bar{x},\bar{y})$ is non-singular, we consider the Fixed Point iteration:

$$x^{(j)} = x^{(j-1)} - (f'_x(\bar{x},\bar{y}))^{-1} f(x^{(j-1)}, y),$$

connected to solving the equation $f(x,y) = 0$. Arguing as above, we can show this iteration generates a sequence $\{x^{(j)}\}_{j=1}\infty$ that converges to $x \in \mathbb{R}^n$ satisfying $f(x,y) = 0$ assuming $f'_x(x,y)$ is Lipschitz continuous for x close to $\bar{x}$ and y close to $\bar{y}$. This defines x as a function $g(y)$ of y for y close to $\bar{y}$. We have now proved the (also very famous):

Theorem 54.7 (Implicit Function theorem) *Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ with $f(x,y) \in \mathbb{R}^n$ and $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, and assume that $f(\bar{x},\bar{y}) = 0$. Assume that the Jacobian $f'_x(x,y)$ with respect to x is Lipschitz continuous for x close to $\bar{x}$ and y close to $\bar{y}$, and that $f'_x(\bar{x},\bar{y})$ is non-singular. Then for y close to $\bar{y}$, the equation $f(x,y) = 0$ has a unique solution $x = g(y)$. This defines x as a function $g(y)$ of y.*

54.19 Newton's method

We next turn to *Newton's method* for solving an equation $f(x) = 0$ with $f : \mathbb{R}^n \to \mathbb{R}^n$, which reads:

$$x^{(i+1)} = x^{(i)} - f'(x^{(i)})^{-1} f(x^{(i)}), \quad \text{for } i = 0, 1, 2, ..., \tag{54.24}$$

where $x^{(0)}$ is an initial approximation. Newton's method corresponds to Fixed Point iteration for $x = g(x)$ with $g(x) = x - f'(x)^{-1} f(x)$. We shall prove that Newton's method converges quadratically close to a root $\bar{x}$ when $f'(\bar{x})$ is non-singular. The argument is the same is as in the case $n = 1$ considered above. Setting $e^i = \bar{x} - x^{(i)}$, and using $\bar{x} = \bar{x} - f'(x^{(i)})^{-1} f(\bar{x})$ if $f(\bar{x}) = 0$, we have

$$\begin{aligned}\bar{x} - x^{(i+1)} &= \bar{x} - x^{(i)} - f'(x^{(i)})^{-1}(f(\bar{x}) - f(x^{(i)})) \\ &= \bar{x} - x^{(i)} - f'(x^{(i)})^{-1}(f'(x^{(i)}) + E_f(x^{(i)}, \bar{x})) = f'(x^{(i)})^{-1} E_f(x^{(i)}, \bar{x}).\end{aligned}$$

We conclude that

$$\|\bar{x} - x^{(i+1)}\| \leq C \|\bar{x} - x^{(i)}\|^2$$

provided

$$\|f'(x^{(i)})^{-1}\| \leq C,$$

where C is some positive constant. We have proved the following fundamental result:

Theorem 54.8 (Newton's method) *If $\bar{x}$ is a root of $f : \mathbb{R}^n \to \mathbb{R}^n$ such that $f(x)$ is uniformly differentiable with a Lipschitz continuous derivative close to $\bar{x}$ and $f'(\bar{x})$ is non-singular, then Newton's method for solving $f(x) = 0$ converges quadratically if started sufficiently close to $\bar{x}$.*

In concrete implementations of Newton's method we may rewrite (54.24) as

$$\begin{aligned} f'(x^{(i)})z &= -f(x^{(i)}), \\ x^{(i+1)} &= x^{(i)} + z, \end{aligned}$$

where $f'(x^{(i)})z = -f(x^{(i)})$ is a system of equations in z that is solved by Gaussian elimination or by some iterative method.

Example 54.18. We return to the equation (54.21), that is,

$$h(V) = V - k_i f(V) - U(t_{i-1}) = 0.$$

To apply Newton's method to solve the equation $h(V) = 0$, we compute

$$h'(v) = I - k_i f'(v),$$

and conclude that $h'(v)$ will be non-singular at v, if $k_i < \|f'(v)\|^{-1}$. We conclude that Newton's method converges if k_i is sufficiently small and we start close to the root. Again the restriction on the time step is connected to the Lipschitz constant L_f of f, since L_f reflects the size of $\|f'(v)\|$.

54.20 Differentiation under the integral sign

Finally, we show that if the limits of integration of an integral are independent of a variable x_1, then the operation of taking the partial derivative with respect x_1 can be moved past the integral sign. Let then $f : \mathbb{R}^2 \to \mathbb{R}$ be a function of two real variables x_1 and x_2 and consider the integral

$$\int_0^1 f(x_1, x_2)\, dx_2 = g(x_1),$$

which is a function $g(x_1)$ of x_1. We shall now prove that

$$\frac{dg}{dx_1}(\bar{x}_1) = \int_0^1 \frac{\partial f}{\partial x_1}(\bar{x}_1, x_2)\, dx_2, \tag{54.25}$$

which is referred to as "differentiation under the integral sign". The proof starts by writing

$$f(x_1, x_2) = f(\bar{x}_1, x_2) + \frac{\partial f}{\partial x_1}(\bar{x}_1, x_2)(x_1 - \bar{x}_1) + E_f(x_1, \bar{x}_1, x_2),$$

where we assume that

$$|E_f(x_1, \bar{x}_1, x_2)| \leq K_f(\bar{x}_1 - x_1)^2.$$

Taylor's theorem implies this is true provided the second partial derivatives of f are bounded. Integration with respect to x_2 yields

$$\int_0^1 f(x_1, x_2)\, dx_2 = \int_0^1 f(\bar{x}_1, x_2)\, dx_2 + (x_1 - \bar{x}_1) \int_0^1 \frac{\partial f}{\partial x_1}(\bar{x}_1, x_2)\, dx_2 + \int_0^1 E_f(x_1, \bar{x}_1, x_2)\, dx_2.$$

Since

$$|\int_0^1 E_f(x_1, \bar{x}_1, x_2)\, dx_2| \leq K_f(\bar{x}_1 - x_1)^2$$

(54.25) follows after dividing by $(x_1 - \bar{x}_1)$ and taking the limit as x_1 tends to $\bar{x}_1$. We summarize:

Theorem 54.9 (Differentiation under the integral sign) *If the second partial derivatives of $f(x_1, x_2)$ are bounded, then for $x_1 \in \mathbb{R}$,*

$$\frac{d}{dx_1} \int_0^1 f(x_1, x_2)\, dx_2 = \int_0^1 \frac{\partial f}{\partial x_1}(x_1, x_2)\, dx_2 \tag{54.26}$$

EXAMPLE 54.19.

$$\frac{d}{dx} \int_0^1 (1 + xy^2)^{-1}\, dy = \int_0^1 \frac{\partial}{\partial x}(1 + xy^2)^{-1}\, dy = -\int_0^1 \frac{y^2}{(1 + xy^2)^2}\, dy.$$

Chapter 54 Problems

54.1. Sketch the following surfaces in $\mathbb{R}^3$: (a) $\Gamma = \{x : x_3 = x_1^2 + x_2^2\}$, (b) $\Gamma = \{x : x_3 = x_1^2 - x_2^2\}$, (c) $\Gamma = \{x : x_3 = x_1 + x_2^2\}$, (d) $\Gamma = \{x : x_3 = x_1^4 + x_2^6\}$. Determine the tangent planes to the surfaces at different points.

54.2. Determine whether the following functions are Lipschitz continuous or not on $\{x : |x| < 1\}$ and determine Lipschitz constants:

- (a) $f : \mathbb{R}^3 \to \mathbb{R}^3$ where $f(x) = x|x|^2$,

- (b) $f : \mathbb{R}^3 \to \mathbb{R}$ where $f(x) = \sin |x|^2$,
- (c) $f : \mathbb{R}^2 \to \mathbb{R}^3$ where $f(x) = (x_1, x_2, \sin |x|^2)$,
- (d) $f : \mathbb{R}^3 \to \mathbb{R}$ where $f(x) = 1/|x|$,
- (e) $f : \mathbb{R}^3 \to \mathbb{R}^3$ where $f(x) = x \sin(|x|)$, (optional)
- (f) $f : \mathbb{R}^3 \to \mathbb{R}$ where $f(x) = \sin(|x|)/|x|$. (optional)

54.3. For the functions in the previous exercise, determine which are contractions in $\{x : |x| < 1\}$ and find their fixed points (optional).

54.4. Linearize the following functions on $\mathbb{R}^3$ at $x = (1, 2, 3)$:

- (a) $f(x) = |x|^2$,
- (b) $f(x) = \sin(|x|^2)$,
- (c) $f(x) = (|x|^2, \sin(x_2))$,
- (d) $f(x) = (|x|^2, \sin(x_2), x_1 x_2 \cos(x_3))$.

54.5. Compute the determinant of the Jacobian of the following functions: (a) $f(x) = (x_1^3 - 3x_1x_2^2, 3x_1x_2^2 - x_2^3)$, (b) $f(x) = (x_1 e^{x_2} \cos(x_3), x_1 e^{x_2} \sin(x_3), x_1 e^{x_2})$.

54.6. Compute the second order Taylor polynomials at $(0, 0, 0)$ of the following functions $f : \mathbb{R}^3 \to \mathbb{R}$: (a) $f(x) = \sqrt{1 + x_1 + x_2 + x_3}$, (b) $f(x) = (x_1 - 1)x_2x_3$, (c) $f(x) = \sin(\cos(x_1x_2x_3))$, (d) $\exp(-x_1^2 - x_2^2 - x_3^2)$, (e) try to estimate the errors in the approximations in (a)-(d).

54.7. Linearize $f \circ s$, where $f(x) = x_1x_2x_3$ at $t = 1$ with (a) $s(t) = (t, t^2, t^3)$, (b) $s(t) = (\cos(t), \sin(t), t)$, (c) $s(t) = (t, 1, t^{-1})$.

54.8. Evaluate $\int_0^\infty y^n e^{-xy}\,dy$ for $x > 0$ by repeated differentiation with respect to x of $\int_0^\infty e^{-xy}\,dy$.

54.9. Try to minimize the function $u(x) = x_1^2 + x_2^2 + 2x_3^2$ by starting at $x = (1, 1, 1)$ using the method of steepest descent. Seek the largest step length for which the iteration converges.

54.10. Compute the roots of the equation $(x_1^2 - x_2^2 - 3x_1 + x_2 + 4, 2x_1x_2 - 3x_2 - x_1 + 3) = (0, 0)$ using Newton's method.

54.11. Generalize Taylor's theorem for a function $f : \mathbb{R}^n \to \mathbb{R}$ to third order.

54.12. Is the function $f(x_1, x_2) = \frac{x_1^2 - x_2^2}{x_1^2 + x_2^2}$ Lipschitz continuous close to $(0, 0)$?

55
Level curves/surfaces and the gradient

It would make no sense to overload the student with all kinds of little things that might be of occasional use. Instead, it is important that students become familiar with ways to think mathematically, recognize the need for applying mathematical methods to engineering problems, realize that mathematics is a systematic science built on relatively few principles and get a firm grasp for the interrelation between theory, computing and experiment. (E. Kreyszig, in Preface to Advanced Engineering Mathematics, 1993)

55.1 Level curves

A *level curve* of a function $u : \mathbb{R}^2 \to \mathbb{R}$ is a curve $g : [a, b] \to \mathbb{R}^2$ such that

$$u(g(t)) = c \quad \text{for } t \in [a, b], \tag{55.1}$$

where c is a constant. A level curve is also called an *isoline*. The points x on a level curve $x = g(t)$ satisfying (55.1), all have the same function value $u(x) = u(g(t)) = c$. By plotting the level curves or isolines for a collection of different constants c, we get a *level curve plot* or *contour plot* of the function $u(x)$. The level curves are the projections onto $\mathbb{R}^2$ of the intersections of the planes $x_3 = c$ in $\mathbb{R}^3$ with the graph $\{x \in \mathbb{R}^2 : x_3 = u(x_1, x_2), (x_1, x_2) \in \mathbb{R}^2\}$. We illustrate in Fig. 55.2.

EXAMPLE 55.1. The level curves of the function $u(x) = x_1^2 + x_2^2$ are the circles $x_1^2 + x_2^2 = c$ with $c \geq 0$ a constant. The level curves of the

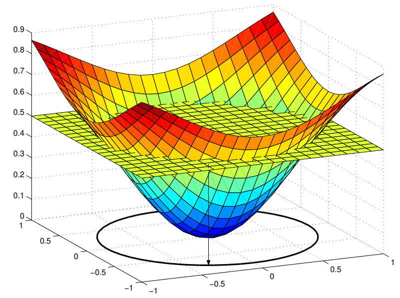


FIGURE 55.1. Projection onto $\mathbb{R}^2$ of the intersection of $x_3 = c$ and $x_3 = u(x_1, x_2)$ (with $u(x_1, x_2) = 1 - \exp(-x_1^2 - x_2^2)$ and $c = .5$) gives a level curve.

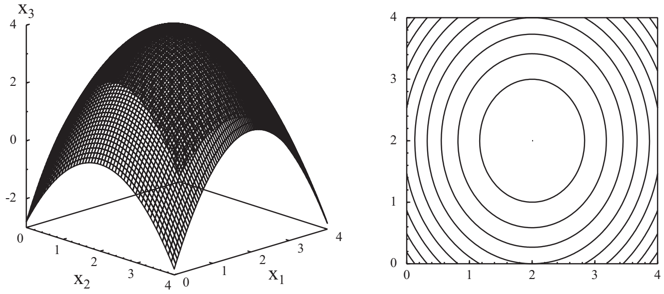


FIGURE 55.2. A plot of a surface and the corresponding contour plot with contour curves shown every .7 units starting at the maximum height of 4.

function $u(x) = 2x_1^2 + x_2^2$ are the ellipses $2x_1^2 + x_2^2 = c$ with $c \geq 0$. The level curves of the function $u(x) = x_1^2 - x_2$ are the parabolas $x_2 = x_1^2 - c$ with c a constant.

EXAMPLE 55.2. A hiking map indicates the level curves of the function $u : \mathbb{R}^2 \to \mathbb{R}$ that gives the height of a point $x \in \mathbb{R}^2$ above a reference level, like the see level. The difference in height between two nearby level curves is typically 10 meters. The change in height between two points can be obtained by counting the number of contour lines intersected by a line joining the two points. This is useful when planinng a hiking trip. Recall Fig. 54.7

A level curve $u(g(t)) = c$ may be thought of as the shore-lines with the sea level equal to c above the reference level.

55.2 Local existence of level curves

The local existence of level curves follows from the following special case of the Implicit Function theorem, where the level curve is given by $t \to (t, g(t))$ or $t \to (g(t), t)$, where $g : \mathbb{R} \to \mathbb{R}$.

Theorem 55.1 *Assume* $u : \mathbb{R}^2 \to \mathbb{R}$ *has continuous partial derivatives and* $u(\bar{x}_1, \bar{x}_2) = c$. *If* $\frac{\partial u}{\partial x_2}(\bar{x}_1, \bar{x}_2) \neq 0$, *then there is a* $\delta > 0$ *such that* $u(x_1, x_2) = c$ *has a unique solution* $x_2 = g(x_1)$ *for* $|x_1 - \bar{x}_1| < \delta$. *If* $\frac{\partial u}{\partial x_1}(\bar{x}_1, \bar{x}_2) \neq 0$, *then there is a* $\delta > 0$ *such that* $u(x_1, x_2) = c$ *has a unique solution* $x_1 = g(x_2)$ *for* $|x_2 - \bar{x}_2| < \delta$.

Notice that if $\frac{\partial u}{\partial x_2}(\bar{x}_1, \bar{x}_2) = 0$, then the level curve is parallel to the x_2-axis, and thus we cannot expect the equation $u(x_1, x_2) = c$ to define x_2 as a function of x_1 (a corresponding function $x_2 = g(x_1)$ would then have infinite slope at $x_1 = \bar{x}_1$).

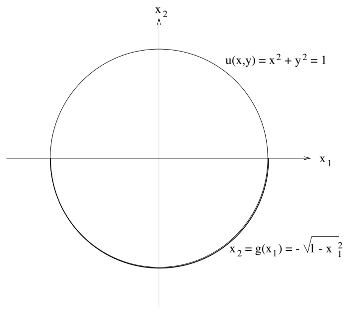


FIGURE 55.3. $x_2 = -\sqrt{1 - x_1^2}$ giving one piece of the level curve $u(x_1, x_2) = x_1^2 + x_2^2 = 1$.

55.3 Level curves and the gradient

Differentiating both sides of (55.1), we get using the Chain rule

$$\frac{d}{dt}u(g(t)) = \nabla u(x) \cdot g'(t) = \frac{\partial u}{\partial x_1}(g(t))g_1'(t) + \frac{\partial u}{\partial x_2}(g(t))g_2'(t) = 0.$$

Since $g'(t) = (g_1'(t), g_2'(t))$ is the direction of the tangent of the curve $g(t)$, this means that the direction $g'(t)$ of a level curve of a function $u : \mathbb{R}^2 \to \mathbb{R}$ is orthogonal to the gradient $\nabla u(g(t))$. Recall that the gradient $\nabla u(x)$ points in the direction of the steepest ascent of the function $u(x)$ at x, and the direction perpendicular to the gradient (the direction of the level curve) is a direction in which u stays constant, see Fig. 55.4. Moving along a level curve the function stays constant, and moving in the direction of the gradient the function increases as quickly as possible!

Since the gradient $\nabla(\bar{x})$ is a normal to the tangent to the level curve through $\bar{x}$, we can write the equation for the tangent to the level curve through $\bar{x}$ in the form $\nabla u(\bar{x}) \cdot (x - \bar{x}) = 0$.

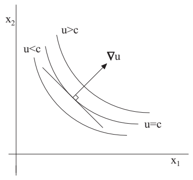


FIGURE 55.4. The gradient of $\nabla u(x)$ of a function $u : \mathbb{R}^2 \to \mathbb{R}$ is orthogonal to the level curve of u through x.

We summarize:

Theorem 55.2 *The gradient $\nabla u(g(t))$ of a function $u : \mathbb{R}^2 \to \mathbb{R}$ is orthogonal to the tangent $g'(t)$ of a level curve $g : I \to \mathbb{R}$. We can write the equation for the tangent to the level curve through $\bar{x}$ in the form $\nabla u(\bar{x})\cdot(x-\bar{x}) = 0$.*

EXAMPLE 55.3. Consider the function $u(x_1, x_2) = x_1^2+x_2^2$ with circular level curves $g(t) = (g_1(t), g_2(t))$ satisfying $g_1^2(t) + g_2^2(t) = c^2$. We have $\nabla u(x) = (2x_1, 2x_2)$ and differentiating $g_1^2(t) + g_2^2(t) = c^2$ with respect to t we get $0 = 2g_1(t)g_1'(t) + 2g_2(t)g_2'(t) = \nabla u(g(t)) \cdot g'(t)$ as expected. Alternatively, parametrizing a level curve $g(t)$ satisfying $g_1^2(t)+g_2^2(t) = c^2$ by $g(t) = c(\cos(t), \sin(t))$, we have $g'(t) = c(-\sin(t), \cos(t)) = (-x_2(t), x_1(t))$ with $x = g(t)$. We check that $\nabla u(g(t))\cdot g'(t) = 2(x_1(t), x_2(t))\times(-x_2(t), x_1(t)) = 0$.

EXAMPLE 55.4. If $u : \mathbb{R}^2 \to \mathbb{R}$ is of the form $u(x_1, x_2) = f(x_1) - x_2$, where $f : \mathbb{R} \to \mathbb{R}$, then $\nabla u(x) = (f'(x_1), -1)$. A level curve $u(g(t)) = c$ can be parametrized by $g(t) = (t, f(t) - c)$, and $g'(t) = (1, f'(t))$. Clearly, $\nabla u(g(t)) \cdot g'(t) = (f'(t), -1) \cdot (1, f'(t)) = 0$.

55.4 Level surfaces

A *level surface* of a function $u : \mathbb{R}^3 \to \mathbb{R}$ is a surface $g : Q \to \mathbb{R}^3$, where Q is a subset of $\mathbb{R}^2$, such that

$$u(g(y)) = c \quad \text{for } y \in Q, \tag{55.2}$$

where c is a constant. A level surface is also called an *isosurface*. The points on a level surface $g(t)$ satisfying (55.1) all have the same function value $u(g(y)) = c$.

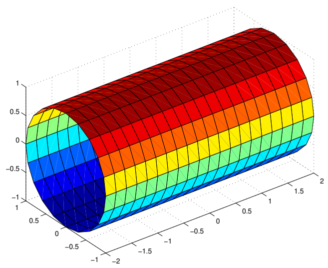


FIGURE 55.5. A piece of the level surface $u(x_1, x_2, x_3) = x_1^2 + x_3^2 = 1$.

55.5 Local existence of level surfaces

The local existence of level surfaces follows from the following special case of the Implicit Function theorem. We find that the level surface is parameterized as $g(y_1, y_2) = (y_1, y_2, f(y_1, y_2))$, $g(y_1, y_3) = (y_1, f(y_1, y_2), y_3)$ or $g(y_2, y_3) = (f(y_2, y_3), y_2, y_3)$ with some function $f : \mathbb{R}^2 \to \mathbb{R}$, depending on which partial derivative is non-zero.

Theorem 55.3 *Assume $u : \mathbb{R}^3 \to \mathbb{R}$ has continuous partial derivatives and $u(\bar{x}_1, \bar{x}_2, \bar{x}_3) = c$, where c is a constant. If $\partial u/\partial x_3 \neq 0$, then there is a $\delta > 0$ such that $u(x_1, x_2, x_3) = c$ has a unique solution $x_3 = f(x_1, x_2)$ for $\|(x_1, x_2) - (\bar{x}_1, \bar{x}_2)\| < \delta$. If $\partial u/\partial x_2 \neq 0$, then there is a $\delta > 0$ such that $u(x_1, x_2, x_3) = c$ has a unique solution $x_2 = g(x_1, x_3)$ for $\|(x_1, x_3) - (\bar{x}_1, \bar{x}_3)\| < \delta$. If $\partial u/\partial x_1 \neq 0$, then there is a $\delta > 0$ such that $u(x_1, x_2, x_3) = c$ has a unique solution $x_1 = g(x_2, x_3)$ for $\|(x_2, x_3) - (\bar{x}_2, \bar{x}_3)\| < \delta$.*

55.6 Level surfaces and the gradient

Differentiating both sides of (55.2) with respect to y_1 and y_2, where $y = (y_1, y_2)$, we get using the Chain rule

$$\frac{\partial}{\partial y_i} u(g(y)) = \nabla u(g(y)) \cdot g'_{,i}(y) = 0, \quad i = 1, 2,$$

where we use the notation

$$g'_{,i}(y) = \frac{\partial}{\partial y_i} g(y).$$

We use the comma in $g'_{,i}$ to indicate differentiation with respect to x_i, while g_i will denote component i of $g = (g_1, g_2, g_3)$. We recall that the tangent plane (linearization) of $g(y)$ at $\bar{x} = g(\bar{y})$ is given by $(y_1, y_2) \to g(\bar{y}) + (y_1 - \bar{y}_1)g'_{,1}(\bar{y}) + (y_2 - \bar{y}_2)g'_{,2}(\bar{y})$, and we conclude that $\nabla u(g(\bar{y}))$ is orthogonal to the tangent plane of the level surface through $\bar{x} = g(\bar{y})$. We say that $\nabla u(g(\bar{y}))$ is *orthogonal to the level surface* $u(x) = c$ through $\bar{x} = g(\bar{y})$, or that $\nabla u(g(\bar{y}))$ is a *normal to the level surface* $u(x) = c$ at $\bar{x} = g(\bar{y})$, see Fig. 55.6. Since $\nabla u(\bar{x})$ thus is a normal to the tangent plane at x, the equation for the tangent plane to a level surface through $\bar{x}$ can also be written $\nabla u(\bar{x}) \cdot (x - \bar{x}) = 0$.

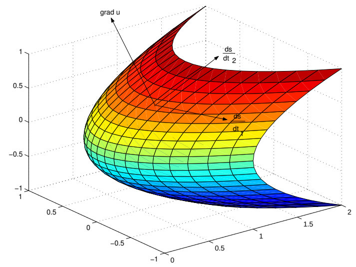


FIGURE 55.6. The gradient $\nabla u(x) = (2x_1, -1, 2x_3)$ of $u(x_1, x_2, x_3) = x_1^2 + x_3^2 - x_2$ is orthogonal to a level surface $(x_1, x_3) \to g(x_1, x_3) = (x_1, x_1^2 + x_3^2 + c, x_3)$ since $g'_1 = (1, 2x_1, 0)$ and $g'_3 = (0, 2x_3, 1)$.

We summarize:

Theorem 55.4 *The gradient $\nabla u(\bar{x})$ of a function $u : \mathbb{R}^3 \to \mathbb{R}$, is orthogonal to the tangent plane $(y_1, y_2) \to g(\bar{y}) + (y_1 - \bar{y}_1)g'_{,1}(\bar{y}) + (y_2 - \bar{y}_2)g'_{,2}(\bar{y})$ of a level surface $y \to x = g(y)$, where $\bar{x} = g(\bar{y})$. The equation for the tangent plane of a level surface through $\bar{x}$ can also be written $\nabla u(\bar{x}) \cdot (x - \bar{x}) = 0$.*

EXAMPLE 55.5. Consider the function $u(x) = x_1^2 + x_2^2 + x_3^2$ with the level surfaces $g(y)$ satsisfying $g_1^2(y) + g_2^2(y) + s_3^2(y) = c^2$ representing spheres centered at the origin with radii c. The gradient $\nabla(x) = 2x$ is evidently orthogonal to a tangent plane of a level surface at x.

EXAMPLE 55.6. If $u : \mathbb{R}^3 \to \mathbb{R}$ is of the form $u(x_1, x_2, x_3) = f(x_1, x_2) - x_3$, where $f : \mathbb{R} \to \mathbb{R}$, then $\nabla u(x) = (f'_{,1}(x_1, x_2), f'_{,2}(x_1, x_2), -1)$. A level surface $u(g(y)) = c$ can be parametrized by $g(y) = (y_1, y_2, f(y_1, y_2) - c)$, and $g'_1(y) = (1, 0, f'_{,1}(y))$ and $g'_2(y) = (0, 1, f'_{,2}(y))$. Clearly, $\nabla u(g(y)) \cdot g'_{,i}(y) = 0$ for $i = 1, 2$.

Chapter 55 Problems

55.1. Sketch the following surfaces in $\mathbb{R}^3$: (a) $\Gamma = \{x : x_1^2 + x_2^2 = x_3\}$, (b) $\Gamma = \{x : x_1^2 + x_2^2 = x_3^2\}$, (c) $\Gamma = \{x : x_1^2 + x_2^2 = -x_3^2\}$, (d) $\Gamma = \{x : x_1^2 + 2x_2^2 + 3x_3^2 = 6\}$. Determine the tangent planes to the surfaces at various points.

55.2. Find parametrization of the curves for the intersections of the surfaces in the previous exercise with the plane $x_3 = 1$.

55.3. Show that the surface $\Gamma = \{x : x_1^2 + 2x_2^2 + 3x_3^2 + x_1 x_3^3 = 7\}$ can be expressed in the form $x_3 = g(x_1, x_2)$ close to $(1, 1, 1)$.

55.4. Compute the gradients of the following functions $f : \mathbb{R}^3 \to \mathbb{R}$: (a) $f(x) = x_1^n(x_2^n + x_3^n)$, (b) $f(x) = |x|$, (c) $f(x) = |x|^2$, (d) $f(x) = 1/|x|$, (e) $f(x) = \exp(x_1 x_2 x_3)$.

55.5. For each of the functions in the previous exercise, determine the equation for the tangent plane to the level surface $f(x) = f(1, 1, 1)$ at $x = (1, 1, 1)$.

55.6. Determine the equation for the tangent plane at $x = (1, 2, 3)$ for the following surfaces: (a) $x_3 = \frac{3}{2} x_1 x_2$, (b) $x_1^2 + x_2^2 + x_3^2 = 14$, (c) $x_2 = \sin(2\pi x_1) + 2\cos(2\pi x_3)$.

55.7. Determine the tangent plane and normal vector to the ellipse $x_1^2 + 3x_2^2 = 10$ at $x = (1, \sqrt{3})$.

55.8. Let $f : Q \to \mathbb{R}$, where $Q = [0, 1] \times [0, 1]$ is the unit square, satisfy $f(x) = 0$ for x on the boundary of Q. Prove under convenient assumptions that there is a point $y \in Q$ such that $\nabla f(y) = 0$.

56

Linearization and stability of initial value problems

The logos of somewome to that base anything, when most characteristically mantissa minus, comes to nullum in the endth: orso, here is nowet badder than the sin of Aha with his cosin Lil, verswaysed on coversvised, and all that's consecants and cotangincies... (Finnegans Wake, James Joyce)

56.1 Introduction

We continue the study of the general initial value problem (40.1), now focussing on the *stability* of solutions, which is a measure of the *sensitivity of solutions to perturbations in given data*. This is a fundamentally important aspect of the behavior of solutions, which we touched upon in Chapter *The general initial value problem*, and which we now consider more closely.

We consider an autonomous problem of the form

$$\dot{u}(t) = f(u(t)) \quad \text{for } 0 < t \leq T,\, u(0) = u^0, \tag{56.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}^d$ is a given bounded Lipschitz continuous function, $u^0 \in \mathbb{R}^d$ is a given initial value, and we seek a solution $u : [0, T] \to \mathbb{R}^d$, where we think of $[0, T]$ as a given time interval. To study the stability of a given solution $u(t)$ to small perturbations in given data, e.g. in the given initial data u^0, we will consider an associated *linearized problem* that arises upon linearizing the function $v \to f(v)$ around the solution $u(t)$.

56.2 Stationary solutions

We consider first the simplest case of a *stationary solution* $u(t) = \bar{u}$ for $0 \leq t \leq T$, that is a solution $u(t)$ of (56.1) that is independent of time t. Since $\dot{u}(t) = 0$ if $u(t)$ is independent of time, $u(t) = \bar{u}$ is a stationary solution if $f(\bar{u}) = 0$ and $u^0 = \bar{u}$, where $\bar{u} = (\bar{u}_1, ..., \bar{u}_d) \in \mathbb{R}^d$. The equation $f(\bar{u}) = 0$ corresponds to a system of d equations $f_i(\bar{u}_1, ..., \bar{u}_d) = 0$, $i = 1, ..., d$, in d unknowns $\bar{u}_1, ..., \bar{u}_d$, where the f_i are the components of f. We studied such systems of equations in Chapter *Vector-valued functions of several real variables*. Here, we assume the existence of a stationary solution $u(t) = \bar{u}$ so that $\bar{u} \in \mathbb{R}^d$ satisfies the equation $f(\bar{u}) = 0$. In general, there may be several roots $\bar{u}$ of the equation $f(v) = 0$ and thus there may be several stationary solutions. We also refer to a stationary solution $u(t) = \bar{u}$ as an *equilibrium solution*.

EXAMPLE 56.1. The stationary solutions $\bar{u}$ of the Crash model

$$\begin{cases} \dot{u}_1 + \nu u_1 - \kappa u_1 u_2 = \nu & t > 0, \\ \dot{u}_2 + 2\nu u_2 - \nu u_2 u_1 = 0 & t > 0, \end{cases} \tag{56.2}$$

of the form $\dot{u} = f(u)$ with $f(u) = (-\nu u_1 + \kappa u_1 u_2 + \nu, -2\nu u_2 + \nu u_2 u_1)$, are $\bar{u} = (1, 0)$ and $\bar{u} = (2, \frac{\nu}{\kappa})$.

56.3 Linearization at a stationary solution

We shall now study perturbations of a given stationary solution under small perturbations of initial data. We thus assume $f(\bar{u}) = 0$ and denote the corresponding equilibrium solution by $\bar{u}(t)$ for $t > 0$, that is $\bar{u}(t) = \bar{u}$ for $t > 0$. We consider the initial value problem (56.1) with $u^0 = \bar{u} + \varphi^0$, where $\varphi^0 \in \mathbb{R}^d$ is a given small perturbation of the initial data $\bar{u}$. We denote the corresponding solution by $u(t)$ and focus attention on the corresponding perturbation in the solution, that is $\psi(t) = u(t) - \bar{u}(t) = u(t) - \bar{u}$. We want to derive a differential equation for the perturbation $\psi(t)$, and to this end we linearize f at $\bar{u}$ and write

$$f(u(t)) = f(\bar{u} + \psi(t)) = f(\bar{u}) + f'(\bar{u})\psi(t) + e(t),$$

where $f'(\bar{u})$ is the Jacobian of $f : \mathbb{R}^d \to \mathbb{R}^d$ at $\bar{u}$ and the error term $e(t)$ is quadratic in $\psi(t)$ (and thus is very small if $\psi(t)$ is small). Since $f(\bar{u}) = 0$ and $u(t)$ satisfies (56.1), we have

$$\dot{\psi}(t) = \frac{d}{dt}(\bar{u} + \psi(t))) = f(u(t)) = f'(\bar{u})\psi(t) + e(t).$$

Neglecting the quadratic term $e(t)$, we are led to a linear initial value problem,

$$\dot{\varphi}(t) = f'(\bar{u})\varphi(t) \quad \text{for } t > 0, \quad \varphi(0) = \varphi^0, \tag{56.3}$$

where $\varphi(t)$ is an approximation of the perturbation $\psi(t) = u(t) - \bar{u}$ up to a second order term. We refer to (56.3) as the *linearized problem* associated to the stationary solution $\bar{u}$ of (56.1). Since $f'(\bar{u})$ is a constant $d \times d$ matrix, we can express the solution to (56.3) using the matrix exponential as

$$\varphi(t) = \exp(tA)\varphi^0 \quad \text{for } 0 < t \leq T, \tag{56.4}$$

where $A = f'(\bar{u})$. We thus have a formula that describes the evolution of perturbation $\varphi(t)$ starting from an initial perturbation $\varphi(0) = \varphi^0$. Depending on the nature of the matrix $\exp(tA)$, the perturbation may increase or decrease with time, reflecting a stronger or lesser sensitivity of the solution $u(t)$ to perturbations in initial data and therefore different stability features of the given problem.

We know that if A is diagonalizable, so that $A = B\Lambda B^{-1}$ where B is a non-singular $d \times d$ matrix and Λ is a diagonal matrix with the eigenvalues $\lambda_1, ..., \lambda_d$ of A on the diagonal, then

$$\varphi(t) = B\exp(t\Lambda)B^{-1}\varphi^0 \quad \text{for } t \geq 0. \tag{56.5}$$

We see that each component of $\varphi(t)$ is a linear combination of $\exp(t\lambda_1)$,...,$\exp(t\lambda_d)$ and the sign of the real part Re λ_i of λ_i determines if the corresponding term grows or decays exponentially. If some Re $\lambda_i > 0$, then we have exponential growth of certain perturbations, which indicates that the corresponding stationary solution $\bar{u}$ is *unstable*. On the other hand, if all Re $\lambda_i \leq 0$, then we would expect $\bar{u}$ to be *stable*.

These considerations are qualitative in nature, and to be more precise we should base judgements of stability or instability on quantitative estimates of perturbation growth. In the diagonalizable case, (56.5) implies in the Euclidean vector and matrix norms that

$$\|\varphi(t)\| \leq \|B\|\|B^{-1}\| \max_{i=1,\dots,d} \exp(t\lambda_i)\|\varphi^0\|. \tag{56.6}$$

We see that the maximal perturbation growth is governed by the maximal exponential factors $\exp(t\lambda_i)$ as well as the factors $\|B\|$ and $\|B^{-1}\|$ related to the transformation matrix B. If the transformation matrix B is orthogonal, then $\|B\| = \|B^{-1}\| = 1$, and the perturbation growth is governed solely by the exponential factors $\exp(t\lambda_i)$. We give this case special attention:

56.4 Stability analysis in the case $f'(\bar{u})$ is symmetric

If $A = f'(\bar{u})$ is symmetric so that $A = Q\Lambda Q^{-1}$ with Q orthogonal and Λ a diagonal matrix with real diagonal elements λ_i, then

$$\|\varphi(t)\| \leq \max_{i=1,\dots,d} \exp(t\lambda_i)\|\varphi^0\|. \tag{56.7}$$

In particular, if all eigenvalues $\lambda_i \leq 0$ then perturbations $\varphi(t)$ cannot grow with time, and we say that the solution $\bar{u}$ is stable. On the other hand, if some eigenvalue $\lambda_i > 0$ and the corresponding eigenvector is g_i then $\varphi(t) = \exp(t\lambda_i)g_i$ solves the linearized initial value problem (56.3) with $\varphi^0 = g_i$, and evidently the particular perturbation $\varphi(t)$ grows exponentially. We then say that the solution $\bar{u}$ is *unstable*. Of course, the size of the positive eigenvalues influence the perturbation growth, so that if $\lambda_i > 0$ is small, then then growth is slow and the instability is mild. Likewise, if λ_i is small negative, then the exponential decay is slow.

56.5 Stability factors

We may express the stability features of a particular perturbation φ^0 through a *stability factor* $S(T, \varphi_0)$ defined as follows:

$$S(T, \varphi_0) = \max_{0 \leq t \leq T} \frac{\|\varphi(t)\|}{\|\varphi^0\|}.$$

where $\varphi(t)$ solves the linearized problem (56.3) with initial data φ^0. The stability factor $S(T, \varphi_0)$ measures the maximal growth of the norm of $\varphi(t)$ over the time interval $[0, T]$ versus the norm of the initial value φ_0.

We can now seek to capture the overall stability features of a stationary solution $\bar{u}$ by maximization over all different perturbations:

$$S(T) = \max_{\varphi^0 \neq 0} S(T, \varphi_0).$$

If the stability factor $S(T)$ is large, then some perturbations grow very much over the time interval $[0, T]$, which indicates a strong sensitivity to perturbations or *instability*. On the other hand, if $S(T)$ is of moderate size then the perturbation growth is moderate, which signifies *stability*. Using the Euclidean matrix norm, we can also express $S(T)$ as

$$S(T) = \max_{0 \leq t \leq T} \|\exp(tA)\|.$$

EXAMPLE 56.2. If $A = f'(\bar{u})$ is symmetric with eigenvalues $\lambda_1, ..., \lambda_d$, then

$$S(T) = \max_{i=1,\ldots,d} \max_{0 \leq t \leq T} \exp(t\lambda_i).$$

In particular, if all $\lambda_i \leq 0$, then $S(T) = 1$.

EXAMPLE 56.3.

The initial value problem for a pendulum takes the form

$$\begin{aligned} \dot{u}_1 &= u_2, \quad \dot{u}_2 = -\sin(u_1) \quad \text{for } t > 0, \\ u_1(0) &= u_{01},\, u_2(0) = u_{02}, \end{aligned}$$

corresponding to $f(u) = (u_2, -\sin(u_1)$ and the equilibrium solutions are $\bar{u} = (0,0)$ and $\bar{u} = (\pi, 0)$. We have

$$f'(\bar{u}) = \begin{pmatrix} 0 & 1 \\ -\cos(\bar{u}_1) & 0 \end{pmatrix},$$

and the linearized problem at $\bar{u} = (0,0)$ thus takes the form

$$\dot{\varphi}(t) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \varphi(t) \equiv A_0\varphi(t) \quad \text{for } t > 0, \quad \varphi(0) = \varphi^0,$$

with solution

$$\varphi_1(t) = \varphi_1^0 \cos(t) + \varphi_2^0 \sin(t), \quad \varphi_2(t) = -\varphi_1^0 \sin(t) + \varphi_2^0 \cos(t).$$

It follows by a direct computation (or using that $\begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{pmatrix}$ is an orthogonal matrix), that for $t > 0$

$$\|\varphi(t)\|^2 = \|\varphi_0\|^2,$$

and thus the norm $\|\varphi(t)\|$ of a solution $\varphi(t)$ of the linearized equations is constant in time, which means that the stability factor $S(T) = 1$ for all $T > 0$. We conclude that if the norm of a perturbation is small initially, it will stay small for all time. This means that the equilibrium solution $\bar{u} = (0,0)$ is *stable*. More precisely, if the pendulum is perturbed initially a little from its bottom position, the pendulum will oscillate back and forth around the bottom position with constant amplitude. This fits our direct experimental experience of course.

Note that the linearized operator A_0 is non-symmetric; the eigenvalues of A_0 are purely imaginary $\pm i$, which says that $\|\varphi(t)\| = \|\varphi_0\|$, that is a perturbation neither grows nor decays. Another way to derive this fact is to use the fact that A_0 is *antisymmetric*, that is $A_0^\top = -A_0$, which shows that $(A_0\varphi, \varphi) = (\varphi, A_0^\top\varphi) = -(\varphi, A_0\varphi) = -(A_0\varphi, \varphi)$, and thus $(A_0\varphi, \varphi) = 0$, where $(\cdot, \cdot)$ is the $\mathbb{R}^2$ scalar product. It follows from the equation $\dot{\varphi} = A_0\varphi$ upon multiplication by φ that $0 = (\dot{\varphi}, \varphi) = \frac{1}{2}\frac{d}{dt}(\varphi, \varphi) = \frac{1}{2}\frac{d}{dt}\|\varphi\|^2$, which proves that $\|\varphi(t)\|^2 = \|\varphi_0\|^2$.

The linearized problem at $\bar{u} = (\pi, 0)$ reads

$$\dot{\varphi}(t) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \varphi(t) \equiv A_\pi\varphi(t) \quad \text{for } t > 0, \quad \varphi(0) = \varphi^0,$$

with symmetric matrix A_π with eigenvalues ± 1. Since one eigenvalue is positive, the stationary solution $\bar{u} = (\pi, 0)$ is unstable. More precisely, the solution is given by

$$\varphi_1 = \frac{\varphi_1^0}{2}(e^t + e^{-t}) + \frac{\varphi_2^0}{2}(e^t - e^{-t}), \quad \varphi_2 = \frac{\varphi_1^0}{2}(e^t - e^{-t}) + \frac{\varphi_2^0}{2}(e^t + e^{-t}),$$

and due to the exponential factor e^t, perturbations will grow exponentially in time, and thus an initially small perturbation will become large as soon as $t \geq 10$ say. Physically, this means that if the pendulum is perturbed initially a little from its top position, the pendulum will eventually move away from the top position, even if the initial perturbation is very small. This fact of course has direct experimental evidence: to balance a pendulum with the weight in the top position is tricky business. Small perturbations quickly grow to large perturbations and the equilibrium solution $(\pi, 0)$ of the pendulum is unstable.

EXAMPLE 56.4. The linearization of the Crash model (56.2) at the equilibrium solution $\bar{u} = (1, 0)$, takes the form

$$\dot{\varphi}(t) = \begin{pmatrix} -\nu & \kappa \\ 0 & -\nu \end{pmatrix} \varphi(t) \equiv A_{\nu,\kappa}\varphi(t) \quad \text{for } t > 0, \quad \varphi(0) = \varphi^0, \qquad (56.8)$$

The solution is given by $\varphi_2(t) = \varphi_2^0 \exp(-\nu t)$, and $\varphi_1(t) = t\kappa \exp(-\nu t)\varphi_2^0 + \exp(-\nu t)\varphi_1^0$. Clearly, $\varphi_2(t)$ decays monotonically to zero and so does $\varphi_1(t)$ if $\kappa = 0$. But, if $\kappa \neq 0$ then $\varphi_1(t)$ reaches the following value, assuming for simplicity that $\varphi_{01} = 0$,

$$\varphi_1(\nu^{-1}) = \nu^{-1}\kappa \exp(-1)\varphi_2^0,$$

which contains the factor ν^{-1} that is large if ν is small. In other words, the stability factor $S(\nu^{-1}) \sim \nu^{-1}$, which is large if ν is small. Eventually, however, $\varphi_1(t)$ decays to zero. As a result, the equilibrium solution $(1, 0)$ is stable only to small perturbations, since we saw in the Chapter The Crash model that $(1, 0)$ is unstable to perturbations above a certain threshold depending on λ. Note that here the Jacobian $f'(\bar{u}) = A_{\nu,\kappa}$ has a double eigenvalue $-\nu$, but $A_{\nu,\kappa}$ is non-symmetric and the space of eigenvectors is one-dimensional and is spanned by $(1, 0)$. As a result, the term $t\kappa \exp(-\nu t)\varphi_2^0$ with linear growth in t appears; thus in this highly non-symmetric problem (if ν is small), large perturbation growth $\sim \nu^{-1}$ is possible although all eigenvalues are non-positive.

The matrix $A_{\nu,\kappa}$ is an example of a *non-normal* matrix. A non-normal matrix A is a matrix such that $A^\top A \neq AA^\top$. A non-normal matrix may or may not be diagonalizable, and if diagonalizable so that $A = B\Lambda B^{-1}$, we may have $\|B\|$ or $\|B^{-1}$ large, resulting in large stability factors in the corresponding linearized problem, as we just saw (cf. Problem 56.5).

The linearization at the equilibrium solution $\bar{u} = (2, \frac{\nu}{\kappa})$ takes the form

$$\dot{\varphi}(t) = \begin{pmatrix} 0 & 2\kappa \\ \frac{\nu^2}{\kappa} & 0 \end{pmatrix} \varphi(t) \quad \text{for } t > 0, \quad \varphi(0) = \varphi^0. \qquad (56.9)$$

The eigenvalues of the Jacobian are $\pm\sqrt{2}\nu$ and the solution is a linear combination of $\exp(\sqrt{2}\nu t)$ and $\exp(-\sqrt{2}\nu t)$ and thus has one exponentially growing part with growth factor $\exp(\sqrt{2}\nu t)$. The equilibrium solution $u = (2, \frac{\nu}{\kappa})$ is thus unstable.

56.6 Stability of time-dependent solutions

We now seek to extend the scope to linearization and linearized stability for a time-dependent solution $\bar{u}(t)$ of (56.1). We want to study solutions of the form $u(t) = \bar{u}(t) + \psi(t)$, where $\psi(t)$ is a perturbation. Using $\frac{d}{dt}\bar{u} = f(\bar{u})$ and linearizing f at $\bar{u}(t)$, we obtain

$$\frac{d}{dt}(\bar{u} + \psi)(t) = f(\bar{u}(t)) + f'(\bar{u}(t))\psi(t) + e(t),$$

with $e(t)$ quadratic in $\psi(t)$. This leads to the linearized equation

$$\dot{\varphi}(t) = A(t)\varphi(t) \quad \text{for } t > 0, \varphi(0) = \varphi_0, \tag{56.10}$$

where $A(t) = f'(\bar{u}(t))$ is an $d \times d$ matrix that now depends on t if $\bar{u}(t)$ depends on t. We have no analytical solution formula to this general problem and thus although the stability properties of the given solution $\bar{u}(t)$ are expressed through the solutions $\varphi(t)$ of the linearized problem (56.10), it may be difficult to analytically assess these properties. We may define stability factors $S(T, \varphi_0)$ and $S(T)$ just as above, and we may say that a solution $\bar{u}(t)$ is stable if $S(T)$ is moderately large, and unstable if $S(T)$ is large. To determine $S(T)$ in general, we have to use numerical methods and solve (56.10) with different initial data φ^0. We return to the computation of stability factors in the next chapter on adaptive solvers for initial value problems.

56.7 Sum up

The question of stability of solutions to initial value problems is of fundamental importance. We can give an affirmative answer in the case of a stationary solution with corresponding symmetric Jacobian. In this case a positive eigenvalue signifies instability, with the instability increasing with increasing eigenvalue, and all eigenvalues non-positive means stability. The case of an anti-symmetric Jacobian aslo signifies stability with the norm of perturbations being constant in time. If the Jacobian is non-normal we have to watch out and remember that just looking at the sign of the real part of eigenvalues may be misleading: in the non-normal case algebraic growth may in fact dominate slow exponential decay for finite time. In these cases and also for time-dependent solutions, an analytical stability analysis may be out of reach and the desired information about stability may be obtained by numerical solution of the associated linearized problem.

Chapter 56 Problems

56.1. Determine the stationary solutions to the system

$$\begin{aligned}\dot{u}_1 &= u_2(1-u_1^2),\\ \dot{u}_2 &= 2-u_1u_2,\end{aligned}$$

and study the stability of these solutions.

56.2. Determine the stationary solutions to the following system (Minea's equation) for different values of $\delta > 0$ and γ,

$$\begin{aligned}\dot{u}_1 &= -u_1-\delta(u_2^2+u_3^2)+\gamma,\\ \dot{u}_2 &= -u_2-\delta u_1u_2,\\ \dot{u}_3 &= -u_3-\delta u_1u_3,\end{aligned}$$

and study the stability of these solutions.

56.3. Determine the stationary solutions of the system (56.1) with (a) $f(u) = (u_1(1-u_2), u_2(1-u_1))$, (b) $f(u) = (-2(u_1-10)+u_2\exp(u_1), -2u_2-u_2\exp(u_1))$, (c) $f(u) = (u_1+u_1u_2^2+u_1u_3^2, -u_1+u_2-u_2u_3+u_1u_2u_3, u_2+u_3-u_1^2)$, and study the stability of these solutions.

56.4. Determine the stationary solutions of the system (56.1) with (56.1) with (a) $f(u) = (-1001u_1+999u_2, 999u_1-1001u_2)$, (b) $f(u) = (-u_1+3u_2+5u_3, -4u_2+6u_3, u_3)$, (c) $f(u) = (u_2, -u_1-4u_2)$, and study the stability of these solutions.

56.5. Analyze the stability of the following variant of the linearized problem (56.8) with $\epsilon > 0$ small,

$$\dot{\varphi}(t) = \begin{pmatrix}-\nu & \kappa\\ \epsilon & -\nu\end{pmatrix}\varphi(t) \equiv A_{\nu,\kappa,\epsilon}\varphi(t) \quad \text{for } t>0, \quad \varphi(0)=\varphi^0, \tag{56.11}$$

by diagonalizing the matrix $\equiv A_{\nu,\kappa,\epsilon}$. Note that the diagonalization degenerates as ϵ tends to zero (that is, the two eigenvectors become parallel). Check if $A_{\nu,\kappa,\epsilon}$ is a normal or non-normal matrix.

57
Adaptive solvers of initial value prolems

On two occasions I have been asked (by members of Parliament), "Pray, Mr Babbage, if you put into the machine wrong figures, will the right answer come out?". I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question. (Babbage (1792-1871))

57.1 Introduction

In this chapter, we discuss the important issue of *adaptive error control* for numerical methods for initial value problems. This is the subject of automated choice of the time step with the purpose of controlling the numerical error to within a given tolerance level. The basic idea is to combine *feed-back* information from the computation concerning the *residual* of the computed solution and the results of auxiliary computations of *stability factors*. We focus first on on the cG(1) method and then comment on the backward Euler method, also referred to as dG(0), the discontinuous Galerkin method with piecewise constants.

We also discuss the application of cG(1) and dG(0) to a class of so-called *stiff* IVPs typically arising in chemical reaction modeling.

57.2 The cG(1) method

We recall that cG(1), the continuous Galerkin method with polynomials of order 1, for the initial value problem $\dot{u}(t) = f(u(t))$ for $t > 0$, $u(0) = u^0$, with $f : \mathbb{R}^d \to \mathbb{R}^d$, takes the form

$$U(t_n) = U(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(U(t))\,dt, \quad n = 1, 2, ..., \tag{57.1}$$

where $U(t)$ is continuous piecewise linear with nodal values $U(t_n) \in \mathbb{R}^d$ at an increasing sequence of discrete time levels $0 = t_0 < t_1 < ...$, and $U(0) = u^0$. If we evaluate the integral in (57.1) with the midpoint quadrature rule, we obtain the Midpoint method:

$$U(t_n) = U(t_{n-1}) + k_n f\left(\frac{U(t_n) + U(t_{n-1})}{2}\right), \quad n = 1, 2, ..., \tag{57.2}$$

where $k_n = t_n - t_{n-1}$ is the time step. The cG(1)-method is the first in a family of cG(q)-methods with $q = 1, 2, ...,$, where the solution is approximated by continuous piecewise polynomials of order q. The Galerkin "orthogonality" of cG(1) is expressed by the fact that the method can be formulated

$$\int_{t_{n-1}}^{t_n} (\dot{U}(t) - f(U(t))) \cdot v\,dt = 0, \quad n = 1, 2, ..., \tag{57.3}$$

for all $v \in \mathbb{R}^d$. This says that the *residual*

$$R(U(t) = \dot{U}(t) - f(U(t)), \quad t \in [0, T], \tag{57.4}$$

of the continuous piecewise linear approximate solution $U(t)$ is *orthogonal* to the constant functions $v(t) = v \in \mathbb{R}^d$ on each subinterval (t_{n-1}, t_n). The residual $\dot{u}(t) - f(u(t))$ of the exact solution is zero since $\dot{u}(t) = f(u(t))$, while the residual of $R(U(t))$ of the approximate solution $U(t)$ is non-zero in general. Similarly, in cG(q) the residual is orthogonal on (t_{n-1}, t_n) to polynomials of degree $q-1$. Note that (57.1) is a vector equation that reads

$$U_i(t_n) = U_i(t_{n-1}) + \int_{t_{n-1}}^{t_n} f_i(U(t))\,dt, \quad n = 1, 2, ..., i = 1, ..., d,$$

as can be seen from (57.3) upon setting $v = e_i$, $i = 1, ..., d$.

We will now study the problem of *automatic step-size control* with the purpose of keeping the error

$$\|u(T) - U(T)\| \leq TOL,$$

where $T = t_N$ is a final time and TOL is a given tolerance, while using as few time steps as possible. The objective is the same as that of computing an integral over an interval $[0, T]$ using numerical quadrature to a certain tolerance using as few quadrature points as possible. This is exactly the problem we meet in the case of a scalar initial value problem $\dot{u}(t) = f(u(t), t)$ with $f(u(t), t) = f(t)$.

We shall derive an *a posteriori error estimate* in which the final error $\|u(T)-U(T)\|$ is estimated in terms of the residual $R(U(t) = \dot{U}(t)-f(U(t))$ and certain *stability factors* that measure the *accumulation* of the numerical errors introduced in each time step.

The a posteriori error estimate takes the form

$$\|u(T) - U(T)\| \leq S_c(T) \max_{0\leq t\leq T} \|k(t)R(U(t)\|, \tag{57.5}$$

where $k(t) = k_n = t_n - t_{n-1}$ for $t \in [t_{n-1}, t_n)$ and where the stability factor $S_c(T)$ is defined as follows. We consider the linearized problem

$$-\dot{\varphi}(t) = A^\top(t)\varphi(t) \quad \text{for } 0 < t < T,\ \varphi(T) = \varphi^0, \tag{57.6}$$

where

$$A(t) = \int_0^1 f'(su(t) + (1-s)U(t))\, ds.$$

We note that replacing $u(t)$ by $U(t)$ gives the following approximate formula for $A(t)$,

$$A(t) \approx f'(U(t)),$$

assuming $U(t)$ is close to $u(t)$. We conclude that $A(t)$ is close to the Jacobian $f'(u(t))$ of $f(v)$ at $v = u(t)$ if $U(t)$ is a reasonable approximation of $u(t)$ Note that the dual $A^\top(t)$ of $A(t)$ occurs in (57.6). Note further that the linearized dual problem (57.6) runs *backward* in time since the initial value $\varphi(T) = \varphi^0$ is specified at time $t = T$. We are now ready to introduce the following stability factors:

$$\begin{aligned} S_d(T) &= \max_{\varphi^0\in\mathbb{R}^d} \frac{\|\varphi(t)\|}{\|\varphi^0\|}, \\ S_c(T) &= \max_{\varphi^0\in\mathbb{R}^d} \frac{\int_0^t \|\dot{\varphi}(s)\|\, ds}{\|\varphi^0\|}, \end{aligned} \tag{57.7}$$

where φ solves (57.6). We note that the stability factors measure different features of the the dual solution φ. The stability factor $S_d(t)$ measures the maximal perturbation growth over the time interval $[0, T]$. We met this factor in the previous chapter. We shall see that this factor is tailored to measure the effect of an error in the initial data u^0 and the "d" in S_d refers to "data". The stability factor $S_c(t)$ measures the integral of $\|\dot{\varphi}\|$ over $[0, T]$ and is geared to evaluate the error in cG(1) and the "c" in S_c refers to "computation".

We shall give the proof of (57.5) below, first in a very simple case with $n = 1$ and $f(u(t)) = au(t)$ with a a constant and then in the general case.

The proofs are very similar. Before plunging into the proofs, we shall try to digest the a posteriori error estimate, and see how it can be used to design an adaptive algorithm aiming at controlling the final error $\|u(T)-U(T)\|$ on a given tolerance level with as few time steps as possible.

The stability factors $S_c(T)$ and $S_d(T)$ can be computed by numerically solving the linearized dual problem (57.6) with $\varphi^0 = e_i$ for $i = 1, ..., d$. If d is large, then we may reduce the variation of the initial data by limiting the error control to certain components only, or by trying to choose φ^0 parallel to $u(T)-U(T)$, which we approximate as $U_h(T)-U_H(T)$ with $U_h(T)$ and $U_H(T)$ being approximations computed with two different tolerances.

57.3 Adaptive time step control for cG(1)

We recall the basic error estimate (57.5):

$$\|u(T)-U(T)\| \leq S_c(T) \max_{0\leq t\leq T} \|k(t))R(t)\|, \tag{57.8}$$

where $R(t) = \dot{U}(t) - f(U(t))$ and we assume that the stability factor $S(T)$ has been computed or estimated. We will return to this issue below. To achieve $\|u(T)-U(T)\| \leq TOL$, we use (57.5) to choose the time steps $k_n = t_n - t_{n-1}$ so that

$$k(t) = k_n \approx \frac{TOL}{S_c(T)R_n} \quad \text{for } t \in [t_{n-1}, t_n), \tag{57.9}$$

where

$$R_n = \max_{t_{n-1}\leq t\leq t_n} \|\dot{U}(t) - f(U(t))\|$$

is the residual on the time interval $[t_{n-1}, t_n)$. Note that the residual R_n is computable from the computed solution $U(t)$ and if $S_c(T)$ is known, timestepalg gives an equation for the time step $k_n = t_n - t_{n-1}$, where t_{n-1} already known. As with adaptive numerical quadrature, (57.9) yields a nonlinear equation for the time step $k_n = t_n - t_{n-1}$ that we can seek to solve using some form of trial-and-error strategy or by prediction, e.g. replacing R_n by R_{n-1}.

57.4 A posteriori error estimate for cG(1) a linear scalar IVP

We shall now prove an a posteriori error estimate for cG(1) for a a linear scalar IVP of the form

$$\dot{u}(t) = au(t) + f(t) \quad \text{for } t > 0, u(0) = u^0, \tag{57.10}$$

where a is a constant and $f(t)$ is a given function. The analysis is based on representing the error in terms of the solution $\varphi(t)$ of the following dual problem:

$$\begin{cases} -\dot{\varphi} = a\varphi & \text{for } T > t \geq 0, \\ \varphi(T) = e(T), \end{cases} \tag{57.11}$$

where $e = u - U$. Note again that (57.11) runs "backwards" in time starting at time t_N and that the time derivative term $\dot{\varphi}$ has a minus sign. We start from the identity

$$|e(T)|^2 = |e(T)|^2 + \int_0^T e\,(-\dot{\varphi} - a\varphi)\,dt,$$

and integrate by parts to get the following representation of $|e(T)|^2$,

$$|e(T)|^2 = \int_0^T (\dot{e} - ae)\varphi\,dt + e(0)\varphi(0),$$

where we allow $U(0)$ to be different from $u(0)$, corresponding to an error in the initial value $u(0)$. Since u solves the differential equation (57.10), that is $\dot{u} + au = f$, we have

$$\dot{e} - au = \dot{u} - au - \dot{U} + aU = f - \dot{U} + aU,$$

and thus we obtain the following representation of the error $|e(T)|^2$ in terms of the residual $R(U) = \dot{U} - aU - f$ and the dual solution φ ,

$$|e(T)|^2 = \int_0^T (f + aU - \dot{U})\varphi\,dt + e(0)\varphi(0) = -\int_0^{t_N} R(U)\varphi\,dt + e(0)\varphi(0). \tag{57.12}$$

Next, we use the Galerkin orthogonality of cG(1),

$$\int_{t_{n-1}}^{t_n} R(U)\,dt = 0 \quad \text{for } n = 1, 2, ...,$$

to rewrite (57.23) as

$$e(T)^2 = -\int_0^T R(U)(\varphi - \bar{\varphi})\,dt + e(0)\varphi(0), \tag{57.13}$$

where $\bar{\varphi}$ is the mean-value of φ over each time interval, that is

$$\bar{\varphi}(t) = \frac{1}{k_n}\int_{t_{n-1}}^{t_n} \varphi(s)\,ds \quad \text{for } t \in [t_{n-1}, t_n).$$

We shall now use

$$\int_{I_n} |\varphi - \bar{\varphi}|\,dt \leq k_n \int_{I_n} |\dot{\varphi}|\,dt,$$

which follows by integration from the facts that

$$\varphi(t)-\bar{\varphi}(t)=\frac{1}{k_n}\int_{t_{n-1}}^{t_n}(\varphi(t)-\varphi(s))\,ds,$$

and

$$|\varphi(t)-\varphi(s)|\le\int_s^t|\dot{\varphi}(\sigma)|\,d\sigma\le\int_{t_{n-1}}^{t_n}|\dot{\varphi}(\sigma)|\,d\sigma\quad\text{for } s,t\in[t_{n-1},t_n].$$

Thus, (57.13) implies

$$\begin{aligned}|e(T)|^2&\le\sum_{n=1}^N R_n\int_{I_n}|\varphi-\bar{\varphi}|dt+|e(0)||\varphi(0)|\\&\le\sum_{n=1}^N k_n\,R_n\int_{I_n}|\dot{\varphi}|dt+|e(0)||\varphi(0)|,\end{aligned}\tag{57.14}$$

where

$$R_n=\max_{t_{n-1}\le t\le t_n}|R(U(t))|.$$

Bringing out the max of k_nR_n over n, we get

$$|e(T)|^2\le\max_{1\le n\le N}k_nR_n\int_0^{t_N}|\dot{\varphi}|\,dt+|e(0)||\varphi(0)|.$$

Recalling that $\varphi(T)=e(T)$ and using the definitions of $S_c(t_N)$ and $S_d(t_N)$, we get the following final estimate,

$$|e(T)|\le S_c(T)\max_{0\le t\le T}|k(t)R(U(t))|+S_d(T)|e(0)|.$$

The stability factors $S_c(T)$ and $S_d(T)$ measure the effects of the accumulation of error in the approximation. To give the analysis a quantitative meaning, we have to give a quantitative bound of this factor. The following lemma gives an estimate for $S_c(T)$ and $S_d(T)$ in the cases $a\le 0$ and the case $a\ge 0$ with possibly vastly different stability factors. We notice that the solution $\varphi(t)$ of (57.11) is given by the explicit formula

$$\varphi(t)=e(T)\exp(a(T-t)).$$

We see that if $a\le 0$, then the solution $\varphi(t)$ decays as t decreases from T, and the case $a\le 0$ is thus the "stable case". If $a>0$ then the exponential factor $\exp(aT)$ enters, and depending on the size of a this case is "unstable". More precisely, we conclude directly from the explicit solution formula that

Lemma 57.1 *The stability factors $S_c(T)$ and $S_d(T)$ satisfy if $a>0$,*

$$S_d(T)\le\exp(aT),\quad S_c(T)\le\exp(aT),\tag{57.15}$$

and if $a\le 0$, then

$$S_d(T)\le 1,\quad S_c(T)\le 1.\tag{57.16}$$

57.5 A posteriori error estimate for cG(1) for a general IVP

The extension of the a posteriori error analysis to a general IVP $\dot{u} = f(u)$ with $f : \mathbb{R}^d \to \mathbb{R}^d$ goes as follows. We recall that the linearized dual problem takes the form

$$-\dot{\varphi}(t) = A^\top(t)\varphi(t) \quad \text{for } 0 < t < T,\ \varphi(T) = e(T), \tag{57.17}$$

with

$$A(t) = \int_0^1 f'(su(t) + (1-s)U(t))\,ds,$$

where $u(t)$ is the exact solution and $U(t)$ the approximate solution. We now use the fact that

$$\begin{aligned} A(t)e(t) &= \int_0^1 f'(su(t) + (1-s)U(t))e(t)\,ds \\ &= \int_0^1 \frac{d}{ds} f(su(t) + (1-s)U(t))\,ds = f(u(t)) - f(U(t)), \end{aligned} \tag{57.18}$$

where we used the Chain rule and the Fundamental Theorem of Calculus. We start from the identity

$$\|e(T)\|^2 = \|e(T)\|^2 + \int_0^T e \cdot (-\dot{\varphi} - A^\top \varphi)\,dt,$$

and integrate by parts to get the error representation,

$$\|e(T)\|^2 = \int_0^T (\dot{e} - Ae) \cdot \varphi\,dt + e(0) \cdot \varphi(0),$$

where we allow $U(0)$ to be different from $u(0)$, corresponding to an error in the initial value $u(0)$. Since u solves the differential equation $\dot{u} - f(u) = 0$, (57.18) implies

$$\dot{e} - Ae = \dot{u} - f(u) - \dot{U} + f(U) = -\dot{U} + f(U),$$

and thus we obtain the following representation of the error $\|e(T)\|^2$ in terms of the residual $R(U) = \dot{U} - f(U)$ and the dual solution φ,

$$\|e(T)\|^2 = -\int_0^{t_N} R(U)\varphi\,dt + e(0)\varphi(0). \tag{57.19}$$

From this point, the proof proceeds just as in the scalar case considered above and we thus obtain the following a posteriori error estimate

$$\|e(T)\| \le S_c(T) \max_{0 \le t \le T} \|k(t)R(U(t))\| + S_d(T)\|e(0)\|,$$

which can be used a basis for adaptive time step control as described above. The stability factors $S_c(T)$ and $S_d(T)$ may be estimated by solving the dual problem with suitable initial data. The proof of the a posteriori error estimate shows that the stability factors may be defined by

$$\begin{aligned} S_d(T) &= \frac{\|\varphi(t)\|}{\|e(T)\|}, \\ S_c(T) &= \frac{\int_0^t \|\dot{\varphi}(s)\|\, ds}{\|e(T)\|}, \end{aligned} \tag{57.20}$$

where φ solves the linearized dual problem with initial data $\varphi(T) = e(T)$. To compute the stability factors $S_d(T)$ and $S_c(T)$, we may solve the dual problem with some estimation of $e(T)$ obtained by solving the initial value problem with two tolerances and approximating $e(T)$ by the difference of the corresponding approximate solutions.

The size of the stability factors indicate the degree of stability of the solution $u(t)$ being computed. If the stability factors are large, the residuals $R(U(t)$ and $e(0)$ have to be made correspondingly smaller by choosing smaller time steps and the computational problem is more demanding.

57.6 A posteriori error estimates for backward Euler for a general IVP

We now derive an a posteriori error estimate for the backward Euler method for the IVP (56.1):

$$U(t_n) = U(t_{n-1}) + k_n f(U(t_n)), \quad n = 1, 2, ..., N, \quad U(0) = u^0.$$

We associate a function $U(t)$ defined on $[0, T]$ to the function values $U(t_n)$, $n = 0, 1, ..., N$, as follows:

$$U(t) = U(t_n) \quad \text{for } t \in (t_{n-1}, t_n].$$

In other words, $U(t)$ is piecewise constant on $[0, T]$ and takes the value $U(t_n)$ on $(t_{n-1}, t_n]$, and thus takes a jump from the value $U(t_{n-1})$ to the value $U(t_n)$ at the time level t_{n-1}.

We can now write the backward Euler method as,

$$U(t_n) = U(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(U(t))\, dt,$$

or equivalently

$$U(t_n) \cdot v = U(t_{n-1}) \cdot v + \int_{t_{n-1}}^{t_n} f(U(t)) \cdot v\, dt, \tag{57.21}$$

for all $v \in \mathbb{R}^d$. This method os also referred to as dG(0), that is the *discontinuous Galerkin method of order zero, corresponding to approximating the exact solution by a piecewise constant function* $U(t)$ *satisfying the orthogonality condition* (57.21).

We are now ready to derive an a posteriori error estimate following the same strategy as for the cG(1) method. We start from the identity

$$\|e(T)\|^2 = \|e(T)\|^2 + \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} e \cdot (-\dot{\varphi} - A^\top \varphi)\, dt,$$

and integrate by parts on each subinterval (t_{n-1}, t_n) to get the following error representation,

$$\begin{aligned}\|e(T)\|^2 = &\sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} (\dot{e} - Ae) \cdot \varphi\, dt \\ &- \sum_{n=2}^{N-1} (U(t_n) - U(t_{n-1}))\varphi(t_{n-1}),\end{aligned}$$

where the last term results from the jumps of $U(t)$ at the the nodes $t = t_{n-1}$ and we assume $U(0) = u(0)$ for simplicity. Since u solves the differential equation $\dot{u} - f(u) = 0$, (57.18) and the fact that $\dot{U}$ on (t_{n-1}, t_n) imply

$$\dot{e} - Ae = \dot{u} - f(u) - \dot{U} + f(U) = -\dot{U} + f(U) = f(U) \quad \text{on } (t_{n-1}, t_n),$$

and thus we obtain

$$\|e(T)\|^2 = -\sum_{n=2}^{N-1} (U(t_n) - U(t_{n-1}))\varphi(t_{n-1}) + \int_0^{t_N} f(U)\varphi\, dt. \tag{57.22}$$

Using (57.21) with $v = \bar{\varphi}$, the mean value of φ as above, we get

$$\begin{aligned}\|e(T)\|^2 = &-\sum_{n=2}^{N-1} (U(t_n) - U(t_{n-1})) \cdot (\varphi(t_{n-1}) - \bar{\varphi}(t_{n-1})) \\ &+ \sum_{n=1}^{n} \int_{t_{n-1}}^{t_n} f(U)(\varphi - \bar{\varphi})\, dt.\end{aligned}$$

We note that

$$\int_{t_{n-1}}^{t_n} f(U)(\varphi - \bar{\varphi})\, dt = 0,$$

since $f(U(t))$ is constant on $(t_{n-1}, t_n]$, and $\bar{\varphi}$ is the mean value of φ, and thus the error representation takes the final form

$$\|e(T)\|^2 = -\sum_{n=2}^{N-1} (U(t_n) - U(t_{n-1})) \cdot (\varphi(t_{n-1}) - \bar{\varphi}(t_{n-1})). \tag{57.23}$$

Using

$$\|\varphi(t_{n-1}) - \bar{\varphi}(t_{n-1})\| \leq \int_{t_{n-1}}^{t_n} \|\dot{\varphi}(t)\|\, dt,$$

we obtain the following a posteriori error estimate for the backward Euler method ,

$$\|e(T)\| \leq S_c(T) \max_{1\leq n\leq N} \|U(t_n)) - U(t_{n-1}))\|. \tag{57.24}$$

Note the very simple form of this estimate involving the jumps $\|U(t_n)) - U(t_{n-1}))\|$ playing the role the residual. The a posteriori error estimate (57.24) can be used as a basis for an algorithm for adaptive time step control of the following form: for $n = 1, 2, ...$, choose k_n so that

$$\|U(t_n)) - U(t_{n-1}))\| \approx \frac{TOL}{S_c(T)}.$$

57.7 Stiff initial value problems

A *stiff* initial value problem $\dot{u} = f(u)$ may be characterized by the fact that the stability factors $S_d(T)$ and $S_c(T)$ are of moderate size even for large T, while the norm of the linearized operator $f'(u(t))$ is large, that is the Lipschitz constant L_f is very large. Such initial value problems are common for example in models of chemical reaction with reactions on a range of time scales from slow to fast. Typical solutions include so-called *transients* where the fast reactions make the solution change quickly over a short (initial) time interval, after which the fast reactions are "burned out" and the slow reactions make the solution change on a longer time scale.

The prototype of a stiff initial value problem has the form

$$\dot{u} = f(u) \equiv -Au \quad \text{for } t > 0,\, u(t) = u^0 = (u_i^0), \tag{57.25}$$

where A is a constant symmetric positive semidefinite $d \times d$ matrix with non-negative eigenvalues λ_i ranging from zero to large positive values. Accordingly, the norm of the matrix A is large and L_f is large. By diagonalization, we may reduce to the case when A is a diagonal matrix with non-negative diagonal elements λ_i, in which case the solution $u(t) = (u_i(t))$ is given by

$$u_i(t) = \exp(-\lambda_i\, t)u_i^0 \quad \text{for } t > 0, \tag{57.26}$$

with $u^0 = (u_i^0)$. This explicit solution formula shows that a component $u_i(t)$ corresponding to a large positive eigenvalue λ_i decays very quickly to zero, while a component with a small eigenvalue stays almost constant for a long time and eventually decays to zero. The sign of the eigenvalues is evidently crucial: if some λ_i was negative, then the corresponding solution component would explode exponentially more or less quickly depending on the size of λ_i. In particular, (57.26) with the λ_i non-negative implies

$$\|u(t)\| \leq \|u^0\| \quad \text{for } t > 0, \tag{57.27}$$

which indicates a form of stability with stability factor equal to 1 in the sense that the norm of the solution does not increase in time.

The dual problem corresponding to (57.25) takes the form

$$-\dot{\varphi} + A\varphi = 0 \quad \text{for } T > t > 0,\ \varphi(T) = \psi,$$

with ψ given data at time $t = T$. As a counterpart of (57.27), we conclude that $S_d(T) \leq 1$. We can similarly show that $S_c(T)$ grows very slowly with increasing T. We sum up: (57.25) represents a stiff problem; stability factors are of moderate size even for large T while the norm of the (linearized) operator A is large.

From numerical point of view, stiff problems may seem particularly friendly since the stability factors grow very slowly with time, but there is one hook that has attracted a lot of attention in the literature on numerical methods for initial value problems, namely the failure of an explicit method like the forward Euler method. We write this method for the equation $\dot{u} = -Au$ in the form

$$U^n = U^{n-1} - k_n A U^{n-1}$$

with U^n an approximation of $u(t_n)$ and $0 = t_0 < t_1 < \dots$ an increasing sequence of time levels, and $k_n = t_n - t_{n-1}$. If A is diagonal with diagonal elements $\lambda_i \geq 0$, then

$$U_i^n = (1 - k_n\lambda_i)U_i^{n-1}$$

and if λ_i is large positive, then $|1 - k_n\lambda_i|$ may be much larger than 1 unless the time step k_n is sufficiently small ($k_n \leq 2/|\lambda_i|$ for all i) and the the numerical solution will then quickly explode to infinity, while the corresponding exact solution quickly decays to zero. The explicit Euler method will thus give completely wrong results unless sufficiently small time steps are used. This may lead to very inefficient time-stepping since after the transients have died out, the solution may vary only slowly and large time steps would be desirable. We note that the time step limit $k_n \leq 2/|\lambda_i|$ for all i, is set by the largest eigenvalue $\max\lambda_i$, while the time long-time scale is set by the smallest eigenvalue $\min\lambda_i$, so that if the quotient $\max\lambda_i/\min\lambda_i$ is large (which signifies a stiff problem), then explicit Euler would be inefficient outside transients.

On the other hand, the dG(0), or implicit Euler method,

$$U^n + k_n A U^n = U^{n-1}$$

with

$$U_i^n = (1 + k_n\lambda_i)^{-1}U_i^{n-1}$$

will be stable and work very well without step size limitation because $1 + k_n\lambda_i \geq 1$ for all $\lambda_i \leq 0$.

For the cG(1)-method, we will have

$$U_i^n = \frac{1 - k_n\lambda_i}{1 + k_n\lambda_i} U_i^{n-1}$$

and stability prevails because

$$|\frac{1 - k_n\lambda_i}{1 + k_n\lambda_i}| \leq 1$$

for all $\lambda_i \geq 0$.

We conclude that both dG(0) and cG(1) may be used for stiff problems, but both these methods are implicit and require the solution of system of equations at each time step. More precisely, dG(0) for a problem of the form $\dot{u} = f(u)$ takes the form

$$U^n - k_n f(U^n) = U^{n-1}.$$

At each time step we have to solve an equation of the form $v - k_n f(v) = U^{n-1}$ with U^{n-1} given. To this end we may try a damped fixed point iteration in the form

$$v^{(m)} = v^{(m-1)} - \alpha(v^{(m-1)} - k_n f(v^{(m-1)}) - U^{n-1}),$$

with α some suitable matrix (or constant in the simplest case). Choosing $\alpha = I$, and iterating once with $v^0 = 0$ corresponds to the explicit Euler method. Convergence of the fixed point iteration requires that

$$\|I + k_n\alpha f'(v)\| < 1$$

for relevant values of v, which could force α to be small (e.g. in the stiff case with $f'(v)$ having large negative eigenvalues) and result in slow convergence. A first try could be to choose α to be a diagonal matrix with $\alpha_i = (f'_{ii})(v^{m-1}))^{-1}$ (corresponding to *diagonal scaling*) and hope that the number of iterations would not be too large. In some cases more efficient iterative solvers would have to be used.

57.8 On explicit time-stepping for stiff problems

We just learned that explicit time-stepping for stiff problems require small time steps outside transients and thus may be inefficient. We shall now indicate a way to get around this limitation through a process of stabilization, where a large time step is accompanied by a couple of small time steps. The resulting method has similarities with the control system of a modern (unstable) jet fighter like the Swedish JAS Gripen, the flight of which is controlled by quick small flaps of a pair of small extra wings ahead of the main wings, or balancing a stick vertically on the finger tips if we want a more domestic application.

We shall now explain the basic (simple) idea of the stabilization and present some examples, as illustrations of fundamental aspects of adaptive IVP-solvers and stiff problems. Thus to start with, suppose we apply explicit Euler to the scalar problem

$$\begin{aligned} \dot{u}(t) + \lambda u(t) &= 0 \quad \text{for } t > 0. \\ u(0) &= u^0, \end{aligned} \tag{57.28}$$

with $\lambda > 0$ taking first a large time step K satisfying $K\lambda > 2$ and then m small time steps k satisfying $k\lambda < 2$, to get the method

$$U^n = (1 - k\lambda)^m (1 - K\lambda) U^{n-1}, \tag{57.29}$$

altogether corresponding to a time step of size $k_n = K + mk$. Here K gives a large unstable time step with $|1 - K\lambda| > 1$ and k is a small time step with $|1 - k\lambda| < 1$. Defining the polynomial function $p(x) = (1 - \theta x)^m (1 - x)$, where $\theta = \frac{k}{K}$, we can write the method (57.29) in the form

$$U^n = p(K\lambda) U^{n-1}.$$

For stability we need

$$|p(K\lambda)| \leq 1, \quad \text{that is } |1 - k\lambda|^m (K\lambda - 1) \leq 1,$$

or

$$m \geq \frac{\log(K\lambda - 1)}{-\log|1 - k\lambda|} \approx 2\log(K\lambda), \tag{57.30}$$

with $c = k\lambda \approx 1/2$ for definiteness.

We conclude that m may be quite small even if $K\lambda$ is large, since the logarithm grows so slowly, and then only a small fraction of the total time would be spent on stabilizing time-stepping with the small time steps k.

To measure the efficiency gain we introduce

$$\alpha = \frac{1 + m}{K + km} \in (1/K, 1/k),$$

which is the number of time steps per unit interval with stabilized explicit Euler method, and by (57.30)) we have

$$\alpha \approx \frac{1 + 2\log(K\lambda)}{K + \log(K\lambda)/\lambda} \approx 2\lambda \frac{\log(K\lambda)}{K\lambda} \ll 2\lambda, \tag{57.31}$$

for $K\lambda \gg 1$. On the other hand, the number of time steps per unit interval for the usual explicit Euler is

$$\alpha_0 = 1/k = \lambda/2, \tag{57.32}$$

choosing a maximum time step $k = 2/\lambda$.

The cost reduction factor using the stabilized explicit Euler method would thus be

$$\frac{\alpha}{\alpha_0} \approx \frac{4\log(K\lambda)}{K\lambda}$$

which can be quite significant for large values of $K\lambda$.

We now present some examples using an adaptive cg(1) IVP-solver in stabilized explicit form with just a few iterations in each time step, which allows large time steps. In all problems we note the initial transient, where the solution components change quickly, and the oscillating nature of the time step sequence outside the transient with large time steps followed by some small stabilizing time steps.

EXAMPLE 57.1. We apply the indicated method to the scalar problem equation (57.28) with $u^0 = 1$ and $\lambda = 1000$, and display the result in Figure 57.1. The cost reduction factor with comparison to a standard explicit method is large: $\alpha/\alpha_0 \approx 1/310$.

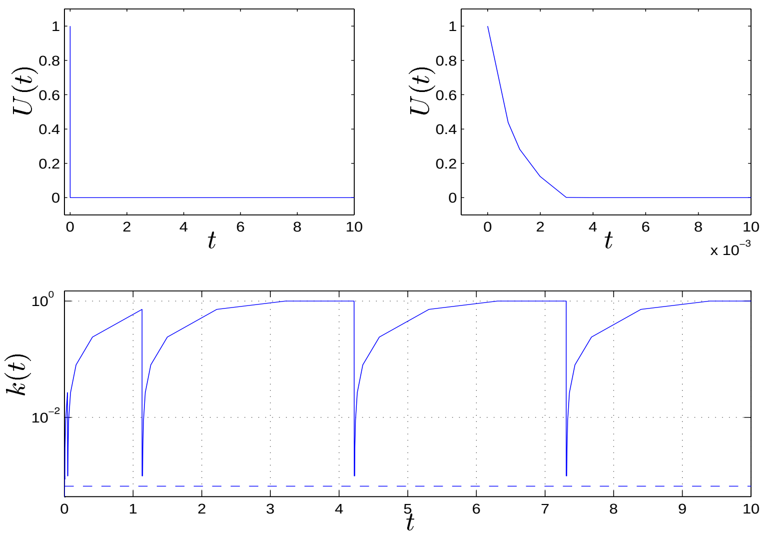


FIGURE 57.1. Solution and time step sequence for eq. (57.28), $\alpha/\alpha_0 \approx 1/310$.

EXAMPLE 57.2. We now consider the 2×2 diagonal system

$$\begin{aligned} \dot{u}(t) + \begin{pmatrix} 100 & 0 \\ 0 & 1000 \end{pmatrix} u(t) = 0 \quad \text{for } t > 0, \\ u(0) = u^0, \end{aligned} \tag{57.33}$$

with $u^0 = (1, 1)$. There are now two eigenmodes with large eigenvalues that have to be stabilized. The cost reduction is $\alpha/\alpha_0 \approx 1/104$.

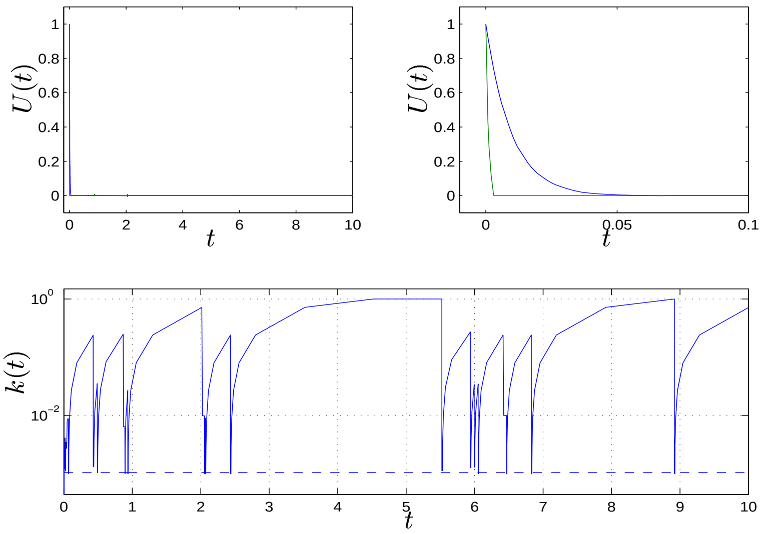


FIGURE 57.2. Solution and time step sequence for eq. (57.33), $\alpha/\alpha_0 \approx 1/104$.

EXAMPLE 57.3. The is the so-called HIRES problem ("High Irradiance RESponse") from plant physiology whcih consists of the following eight equations:

$$\left\{\begin{array}{rcl}\dot{u}_1 & = & -1.71u_1 + 0.43u_2 + 8.32u_3 + 0.0007,\\ \dot{u}_2 & = & 1.71u_1 - 8.75u_2,\\ \dot{u}_3 & = & -10.03u_3 + 0.43u_4 + 0.035u_5,\\ \dot{u}_4 & = & 8.32u_2 + 1.71u_3 - 1.12u_4,\\ \dot{u}_5 & = & -1.745u_5 + 0.43u_6 + 0.43u_7,\\ \dot{u}_6 & = & -280.0u_6u_8 + 0.69u_4 + 1.71u_5 - 0.43u_6 + 0.69u_7,\\ \dot{u}_7 & = & 280.0u_6u_8 - 1.81u_7,\\ \dot{u}_8 & = & -280.0u_6u_8 + 1.81u_7,\end{array}\right. \tag{57.34}$$

together with the initial condition $u^0 = (1.0, 0, 0, 0, 0, 0, 0, 0.0057)$. We present the solution and the time step sequence in Figure 57.3. The cost is now $\alpha \approx 8$ and the cost reduction factor is $\alpha/\alpha_0 \approx 1/33$.

EXAMPLE 57.4. The "Chemical Akzo-Nobel" problem consists of the following six equations:

$$\left\{\begin{array}{rcl}\dot{u}_1 & = & -2r_1 + r_2 - r_3 - r_4,\\ \dot{u}_2 & = & -0.5r_1 - r_4 - 0.5r_5 + F,\\ \dot{u}_3 & = & r_1 - r_2 + r_3,\\ \dot{u}_4 & = & -r_2 + r_3 - 2r_4,\\ \dot{u}_5 & = & r_2 - r_3 + r_5,\\ \dot{u}_6 & = & -r_5,\end{array}\right. \tag{57.35}$$

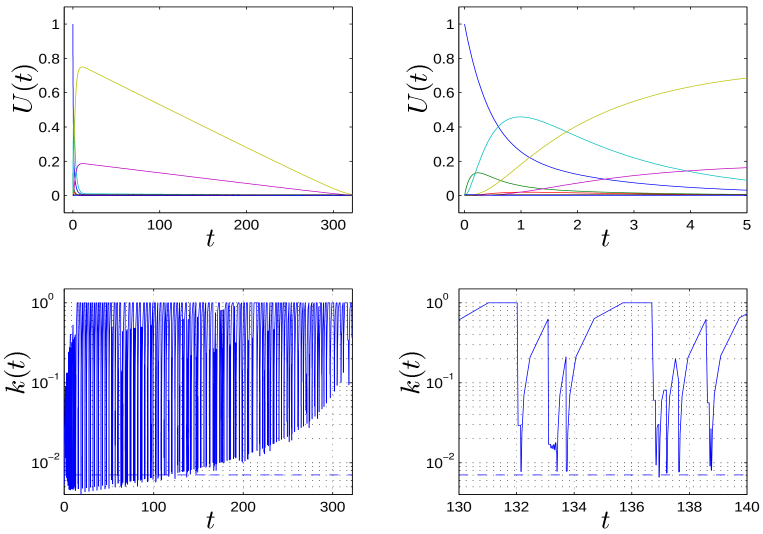


FIGURE 57.3. Solution and time step sequence for eq. (57.34), $\alpha/\alpha_0 \approx 1/33$.

where $F = 3.3 \cdot (0.9/737 - u_2)$ and the reaction rates are given by $r_1 = 18.7 \cdot u_1^4 \sqrt{u_2}$, $r_2 = 0.58 \cdot u_3 u_4$, $r_3 = 0.58/34.4 \cdot u_1 u_5$, $r_4 = 0.09 \cdot u_1 u_4^2$ and $r_5 = 0.42 \cdot u_6^2 \sqrt{u_2}$. We integrate over the interval $[0, 180]$ with initial condition $u^0 = (0.437, 0.00123, 0, 0, 0, 0.367)$. Allowing a maximum time step of $k_{\max} = 1$ (chosen arbitrarily), the cost is only $\alpha \approx 2$ and the cost reduction factor is $\alpha/\alpha_0 \approx 1/9$. The actual gain in a specific situation is determined by the quotient between the large time steps and the small damping time steps, as well as the number of small damping steps that are needed. In this case the number of small damping steps is small, but the large time steps are not very large compared to the small damping steps. The gain is thus determined both by the stiff nature of the problem and the tolerance (or the size of the maximum allowed time step).

EXAMPLE 57.5. We consider now Van der Pol's equation:

$$\ddot{u} + \mu(u^2 - 1)\dot{u} + u = 0,$$

which we write as

$$\left\{ \begin{array}{rcl} \dot{u}_1 & = & u_2, \\ \dot{u}_2 & = & -\mu(u_1^2 - 1)u_2 - u_1. \end{array} \right. \tag{57.36}$$

We take $\mu = 1000$ and solve on the interval $[0, 10]$ with initial condition $u^0 = (2, 0)$. The time step sequence behaves as desired with only a small portion of the time spent on taking small damping steps. The cost is now $\alpha \approx 140$ and the cost reduction factor is $\alpha/\alpha_0 \approx 1/75$.

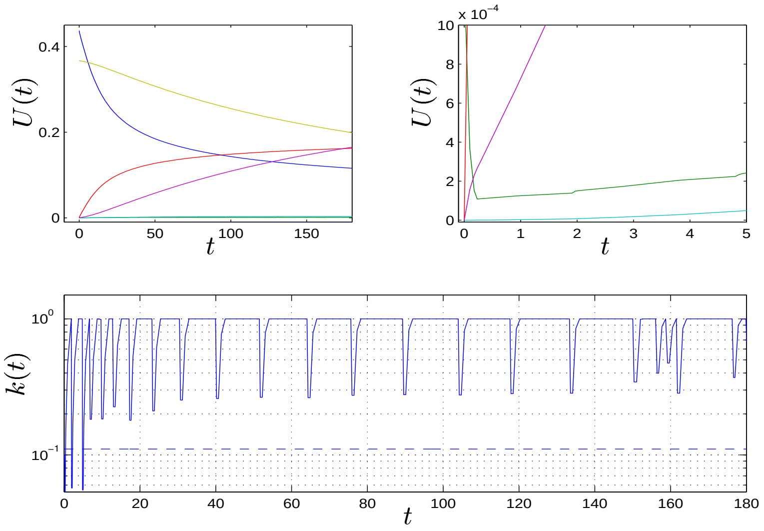


FIGURE 57.4. Solution and time step sequence for eq. (57.35), $\alpha/\alpha_0 \approx 1/9$.

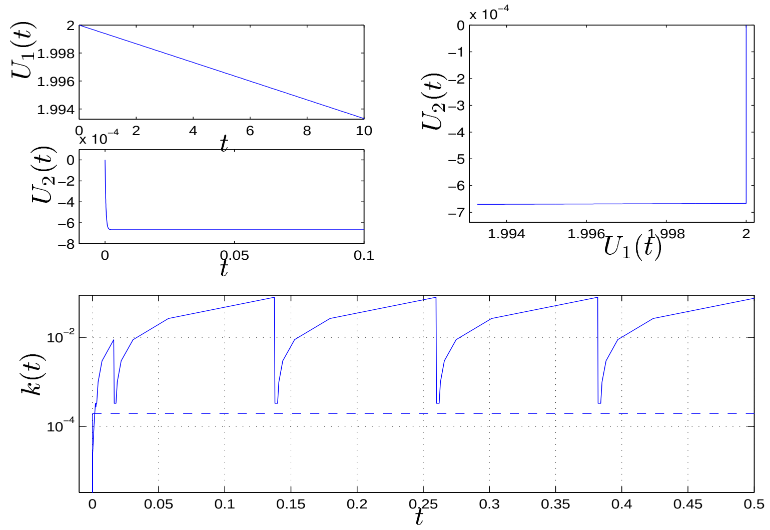


FIGURE 57.5. Solution and time step sequence for eq. (57.36), $\alpha/\alpha_0 \approx 1/75$.

Chapter 57 Problems

57.1. Develop explicit versions of dG(0) and cG(1) based on fixed point iteration at each time step. Show that with diagonal scaling such an explicit method may work very well for some stiff problems.

58
Lorenz and the Essence of Chaos*

> I am convinced that chaos, along with its many associated concepts - strange attractors, basin boundaries, period-doubling bifurcations and the like - can readily be understood and relished by readers who have no special mathematical or other scientific background...
> (E. Lorenz, in Foreword to *The Essence of Chaos*)

58.1 Introduction

On December 29, 1972, the meteorologist Edward Lorenz presented in a session on the Global Atmospheric Research Program at the 139th meeting of the American Association for the Advancement of Science in Washington D.C., a talk with the title *Predictability: Does the Flap of a Butterfly's Wings in Brazil Set off a Tornado in Texas?* The talk by Lorenz with its "Butterfly effect" rocketed to fame a decade later during the development of "Chaos Theory" that became a fashion in mathematics and physics during the 80s, with the pretention of explaining a variety of phenomena from turbulent fluid flow to collapsing stock markets sharing qualities of *unpredictability*. A decade earlier, "Catastrophe Theory" played a similar role, while today very few remember this intriguing subject. Of course, unpredictability or "chaos" is a phenomenon that has long been familiar to mankind. The word "chaos" comes from early Greek cosmology and signifies the complete lack of order of the Universe before the creation of Gaea and Eros (Earth and Desire).

Lorenz' question is connected to the obvious difficulty of making reasonably reliable predictions of the daily weather over longer time than a week. A weather forecast is made by numerically solving an IVP modeling the evolution of the atmosphere, including variables such as temperature, wind speed and pressure. There are many sources of errors in a weather forecast made this way: errors in the initial value, modeling errors and numerical errors, and it seems that these errors are magnified at a rate that limits the predictions, depending on the scale from a few hours in very local models to weeks in global circulation models.

Lorenz' Butterfly analogy indicates that in certain dynamical systems, very small causes may have large effects after some time. We have already met such a problem in the form of a pendulum being released starting from the unstable top position: depending on the initial perturbation the position of the pendulum will be vastly different after some time (one side or the other). In meteorology, this corresponds to a situation where the weather-man can't say if a certain low pressure will take this way or that way, and thus can't be sure if it will rain in Göteborg tomorrow or not. In his book, Lorenz gives other examples of unstable systems such as a pinball machine, where very small changes in the action of the player can change the outcome of the game completely. Of course there are many other examples from real life of "small" causes having large effects, from soccer games to the assassination of Archduke Francis Ferdinand by the Serb nationalist Gavrilo Princip in Sarajevo on June 28 1914, initiating the First World War.

58.2 The Lorenz system

Lorenz formulated an IVP of the form $\dot{u} = f(u)$ with $f : \mathbb{R}^3 \to \mathbb{R}^3$ given by

$$f(u) = (-10u_1 + 10u_2, 28u_1 - u_2 - u_1u_3, -\frac{8}{3}u_3 + u_1u_2),$$

which is the famous *Lorenz system*. Lorenz found that the solution of this system is very sensitive to perturbations. The system has some vague connection to a very simple model for fluid flow and has been given the role of explaining properties of fluid motion, such as turbulence. This was not Lorenz' original idea, who just wanted to make a connection to the apparent unpredictability and supposed sensitivity to perturbations of common meteorological models. If the seemingly very harmless and innocent Lorenz system could have unpredictable solutions, then there should be no surprise that also the weather could be unpredictable.

More precisely, Lorenz found that two solutions of the Lorenz system with very close initial data will stay close for some time but will eventually move apart completely. The Lorenz system is therefore very difficult to solve accurately using a numerical method over times longer than say 30 units. The numerical solution will stay close the the exact solution for some time, but will eventually move apart significantly. Of course there are many IVP:s sharing this property of instability. Even the simple pendulum has this property if the pendulum reaches the top position with small velocity. It is thus remarkable that the Lorenz system seemed to present some kind of surprise to the scientific world. But it did, and it has become quite popular to explain all sorts of phenomena, from turbulence to politics, by referring to the "strange attractor" supposedly being displayed in plots of solutions of the Lorenz system.

The Lorenz system in component form reads:

$$\begin{cases} \dot{u_1} = -10u_1 + 10u_2, \\ \dot{u_2} = 28u_1 - u_2 - u_1u_3, \\ \dot{u_3} = -\frac{8}{3}u_3 + u_1u_2, \\ u_1(0) = u_{01},\ u_2(0) = u_{02},\ u_3(0) = u_{03}, \end{cases} \tag{58.1}$$

and u_0 is a given initial condition. The system (58.1) has three equilibrium points $\bar{u}$ with $f(\bar{u}) = 0$: $\bar{u} = (0,0,0)$ and $\bar{u} = (\pm 6\sqrt{2}, \pm 6\sqrt{2}, 27)$. The equilibrium point $\bar{u} = (0,0,0)$ is unstable with the corresponding Jacobian $f'(\bar{u})$ having one positive (unstable) eigenvalue and two negative (stable) eigenvalues. The equilibrium points $(\pm 6\sqrt{2}, \pm 6\sqrt{2}, 27)$ are slightly unstable with the corresponding Jacobians having one negative (stable) eigenvalue and two eigenvalues with very small positive real part (slightly unstable) and also an imaginary part. More precisely, the eigenvalues at the two non-zero equilibrium points are $\lambda_1 \approx -13.9$ and $\lambda_{2,3} \approx .0939 \pm 10.1i$.

In Fig. 58.1, we present two views of a solution $u(t)$ that starts at $u(0) = (1,0,0)$ computed to time 30 with an error tolerance of $TOL = 0.5$ using an adaptive IVP-solver of the form presented in Chapter *Adaptive IVP-solvers*. We can think of $u(t) = (x(t), y(t), z(t))$ as the position at time t of a particle that moves according to the equation $\dot{u} = f(u)$. In Fig. 58.1 we thus plot the trajectory or path followed by the particle as the particle moves with increasing time. The plotted trajectory is typical: the particle is kicked away from the unstable point $(0,0,0)$ and moves towards one of the non-zero equilibrium points. It then slowly orbits away from that point and at some time decides to cross over towards the other non-zero equilibrium point, again slowly orbiting away from that point and coming back again, orbiting out, crossing over, and so on. This pattern of some orbits around one non-zero equilibrium point followed by a transition to the other non-zero equilibrium point is repeated with a seemingly random number of revolutions around each equilibrium point.

As noted by Lorenz, a close inspection of the trajectory in Fig. 58.1 reveals quite a bit of structure in the behavior of the solution. From the path of the trajectory, it seems that, roughly speaking, there are two flat "lobes" in which the orbits around the non-zero equilibrium points are

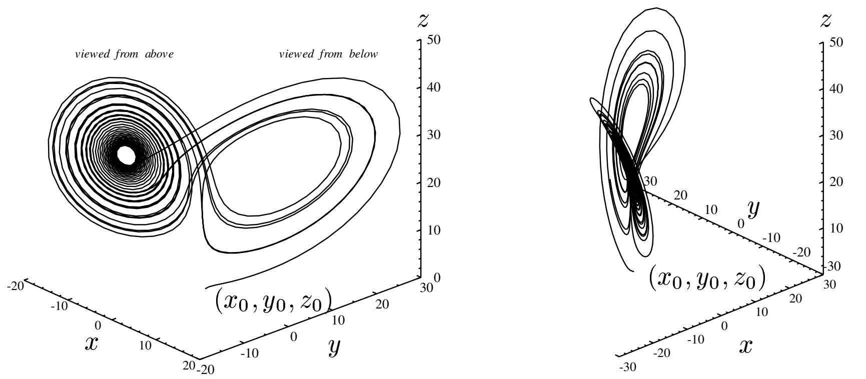


FIGURE 58.1. Two views of a numerical trajectory of the Lorenz system over the time interval $[0, 30]$ starting at $(1, 0, 0)$ computed with absolute error tolerance 0.5.

located. In each lobe, the spiraling segments of the trajectory seem to be grouped in "bands" that are made up of parts of the trajectory that are spiralling out from the equilibrium point and parts of the trajectory that have just crossed over from the other lobe. Only the trajectories in the outer band switch to the other equilibrium point. This causes a sharp separation between trajectories located in the outer band and those located in the next band inside as the trajectories approach the z-axis. We refer to this as *cutting* through the action of a "razor" separating the trajectories in the outer band. The trajectories in the outer band expand in width as they approach the other equilibrium point, with trajectories near the outside of the band ending up nearer to the fixed point. We refer to this as *expansion* and *flipping* respectively. The position of initial approach of the trajectory to an equilibrium point determines the number of orbits the trajectory makes in that lobe before returning to the other equilibrium point. Finally, we see that the orbits in one band come close to the next outer band after one revolution, this repeats with every band of the trajectory, until eventually they all end up in the outer band and leave towards the other equilibrium point. We refer to this as *interlacing*. In short, we can describe the dynamics of the Lorenz system as a never-ending process of cutting, expansion, flipping, and interlacing.

58.3 The accuracy of the computations

The first task is to measure the reliability of the computed error bound based on an a posteriori error estimate of the form presented in Chapter *Adaptive IVP-solvers.* Since we do not have the exact solution, we perform the following experiment: Using the initial data $(0, 1, 0)$, we compute twice using residual tolerances 10^{-5} and 10^{-9} and approximate the error in the less accurate computation by taking the difference between the values of the less accurate and more accurate computations. In Fig. 58.2, we plot the computed error bound and the approximate error. The error bound predicts the size of the error quite well in spite of the sensitivity of the solution to perturbations. Similar results are obtained for a variety of initial data.

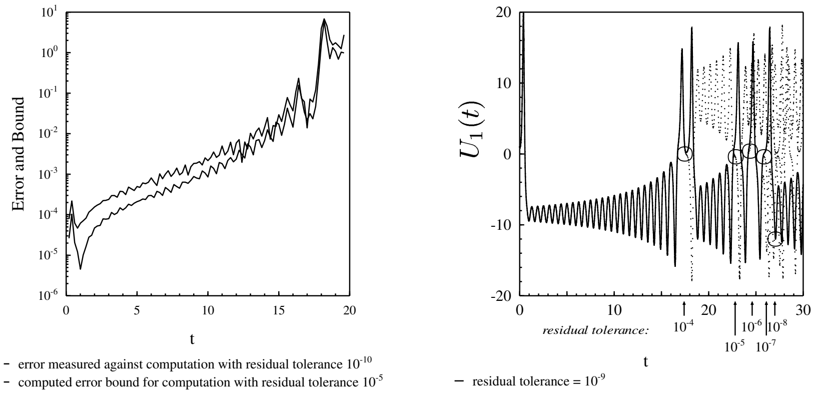


FIGURE 58.2. On the left, we plot the results of the reliability test for the computed error bound with initial condition is $(0, 1, 0)$. On the right, we plot the effect of changing the residual tolerance on the accuracy in $U_1(t)$ component with initial condition is $(1, 0, 0)$.

To give some idea of the behavior of the error control, we plot the step sizes used in a computation with absolute error tolerance 0.75 in Fig. 58.3. The step sizes vary roughly by a factor of 6 over the interval of computation. In Fig. 58.3, we also plot the product of the time step and the residual for this computation. We note that these values are kept within 10% of a constant value. With more computational work, the size of the variations can be reduced, which produces a more smoothly-varying error bound.

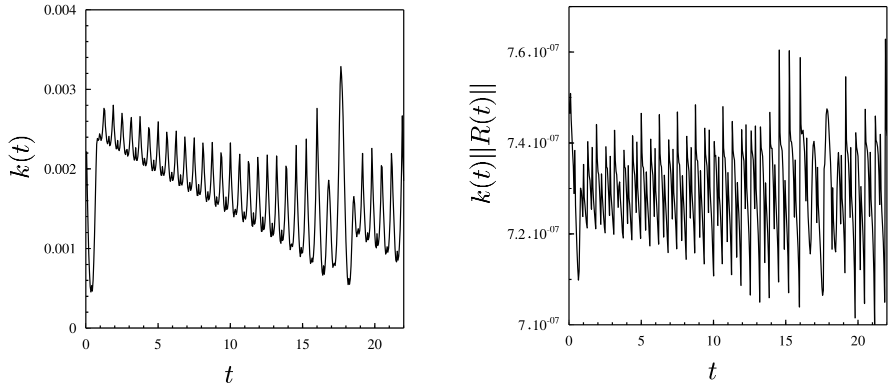


FIGURE 58.3. Time steps and residual $\times$ time step versus time for a computation beginning at $(1, 0, 0)$ with absolute error tolerance 0.75.

58.4 Computability of the Lorenz system

Encouraged by these results we decrease the tolerance or, equivalently, the time step, and try to compute an accurate solution to the Lorenz system on an even longer time interval. Using the cG(1) method as described in Chapter *Adaptive IVP-solvers*, we compute solutions with smaller and smaller time steps, $k = 0.01$, $k = 0.001$ and $k = 0.0001$, and expect to produce more and more accurate solutions. We plot the U_1-component of the solution in Fig. 58.4 where we also indicate the points at which the solutions are no longer accurate. We see that even with $300,000$ time steps the solution is not accurate beyond $t = 26$. Decreasing the time step with a factor 10 or 100 will take us only a little further, but the computation will take 10 or 100 times longer. We conclude that it is difficult to compute the solution to the Lorenz system over long time intervals.

To examine in detail the computability of the Lorenz system we return to the error estimate that we derived for the error $e(t)$ of the cG(1) method:

$$\|e(t)\| \leq S_c(T) \max_{0 \leq t \leq T} \|k(t)R(t)\|. \tag{58.2}$$

Remember that the stability factor $S_c(T)$ for the Lorenz system is defined in terms of the solution to the linearized dual problem as

$$S_c(T) = \max_{\varphi_0 \in \mathbb{R}^3} \frac{\int_0^T \|\dot{\varphi}(t)\| \, dt}{\|\varphi_0\|}.$$

Judging by the error estimate we should be able to reach as far as we want if only the time step $k(t)$ and the residual $R(t)$ are small enough. However,

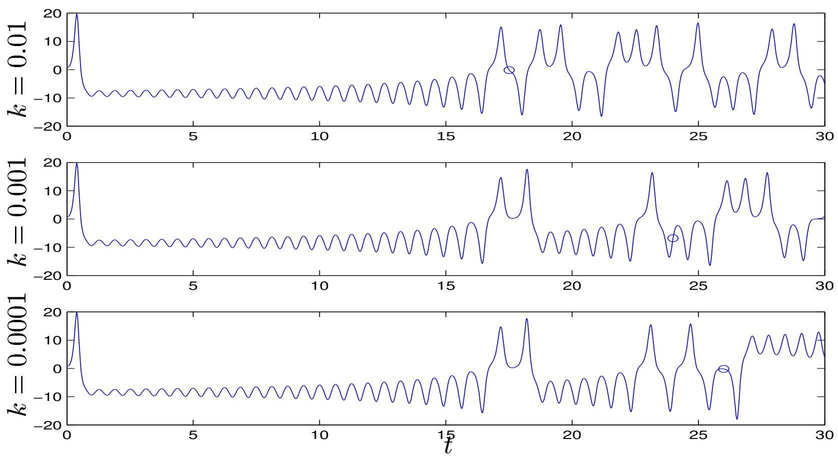


FIGURE 58.4. The U_1-component of the cG(1) solution for different time steps. The small circles indicate the points at which the solutions are no longer accurate.

a little more careful analysis reveals an additional error contribution, which is often ignored. Including also this term into our error estimate, we find:

$$\|e(t)\| \leq S_c(T) \max_{0\leq t\leq T} \|k(t)R(t)\| + S_0(T) \max_{0\leq t\leq T} \epsilon/k(t), \tag{58.3}$$

where ϵ is the *machine precision* of the computer, i.e. the smallest number for which $1+\epsilon \neq 1$ (in computer arithmetic) and $S_0(T)$ is a new stability factor. For a standard computer (in 2002) with so-called double-precision arithmetic, the machine precision is $\epsilon \approx 10^{-16}$. The stability factor $S_0(T)$ is defined in terms of the dual solution as

$$S_0(T) = \max_{\varphi_0 \in \mathbb{R}^3} \frac{\int_0^T \|\varphi(t)\|\, dt}{\|\varphi_0\|}.$$

The additional term in our refined error estimate (58.3) accounts for the round-off error made at every time step in the computation; when the new value $U(t_n)$ for the cG(1) solution is computed in every time step, it is unavoidable that we make a round-off error of size ϵ. As we shall see, it is the second term that sets the limit for the computability of the Lorenz system; the second term in (58.3) can be large even though the first term is small.

The difficulty of computing accurate solutions to the Lorenz system becomes obvious if we plot the size of the stability factors. In Fig. 58.5 we plot the size of the stability factor $S_0(T)$ associated with round-off errors as function of the final time T. Notice the logarithmic scale in this figure. A simple approximation of the growth of this stability factor is

$$S_0(T) \approx 10^{T/3},$$

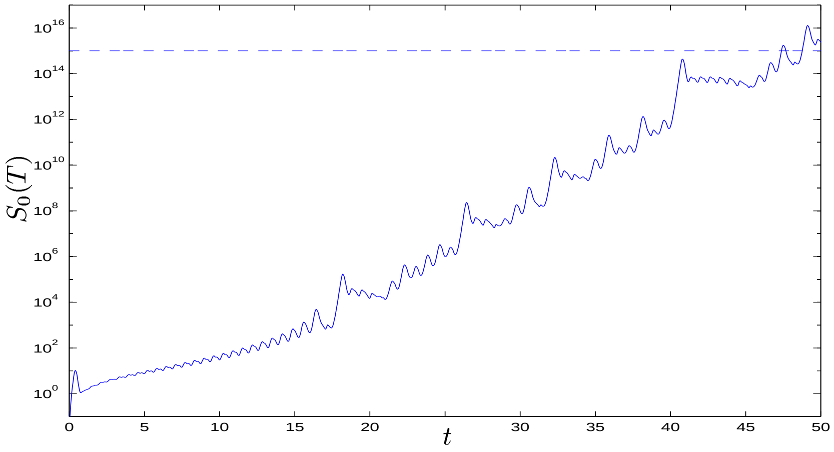


FIGURE 58.5. The growth of the stability factor $S_0(T)$ for the Lorenz problem.

and so the round-off error grows as $E_r = 10^{T/3} \cdot 10^{-16}/k = 10^{T/3-16}/k$. Notice how the error grows larger if we decrease the time step! This is natural (although unusual), since with a smaller time step we will have to take a larger number of time steps and thus make a larger number of round-off errors. To make the influence from round-off errors small we specify a large time step, say $k = 0.1$, for which the round-off error now grows as $10^{T/3-15}$. At time $T = 3 \cdot 15 = 45$ the accumulated round-off error is then $E_r = 1$, which means that we cannot expect to compute much beyond time $T = 45$, since then the round-off error will dominate anyway. Using the cG(1) method, we will not even reach $T = 45$, since we have to use a time step much smaller than $k = 0.1$ (as seen in Fig. 58.4) to make the first term in the error estimate small.

58.5 The Lorenz Challenge

From the previous discussion it is now clear that the mysterious unpredictability and "chaotic" behaviour of the Lorenz system only means that the stability factors grow quickly, making it difficult to compute accurate solutions over long time intervals. The obvious challenge is now, using the method of choice, *to compute an accurate solution to the Lorenz system over as long a time interval* $[0, T]$ *as possible.*

We saw in the previous section that brute force is not the way to go. It is not enough to use a very fast computer with very small and very many time steps. Using the cG(1) method we cannot reach much further than $T = 30$, no matter how small time steps we use since then the accumulated round-off error will grow large. A solution to this problem would be if we could design a method, similar to cG(1), which can be used with larger time steps than what is possible with cG(1). As one can expect, there exist corresponding methods cG(2), cG(3) and so on, which can be used with larger time steps. It can be proved that for these cG(q) methods, the error grows as k^{2q}, i.e. we have so called *a priori* error estimates of the type

$$\|e(T)\| \leq C(T)k^{2q},$$

where $C(T)$ is a constant (unknown!) depending on the exact solution $u(t)$. We say that the cG(q) method is of *order* $2q$. The standard cG(1) method is thus a second order method. (This is in agreement with (58.2) since one factor $k(t)$ is hidden inside $R(t)$.) With a higher order method, i.e. $q > 1$, we can thus obtain a smaller error with the same time step, which makes it possible to compute the solution with larger time steps. This in turn implies that with a higher order method, we can keep the round-off error smaller and thus reach further than what is possible with the cG(1) method. In Fig. 58.6 we plot the U_1-components of solutions to the Lorenz system, computed with time step $k = 0.1$, with a sequence of higher order methods. We see that with high enough order, the solutions agree to a point just beyond $T = 45$ as we predicted; the first term in our error estimate (58.3) has been reduced by increasing the order of the method and so the second term dominates. It is possible to reach beyond time $T = 50$, perhaps to $T = 100$, but to do this we have to go from double-precision arithmetic to quadruple-precision.

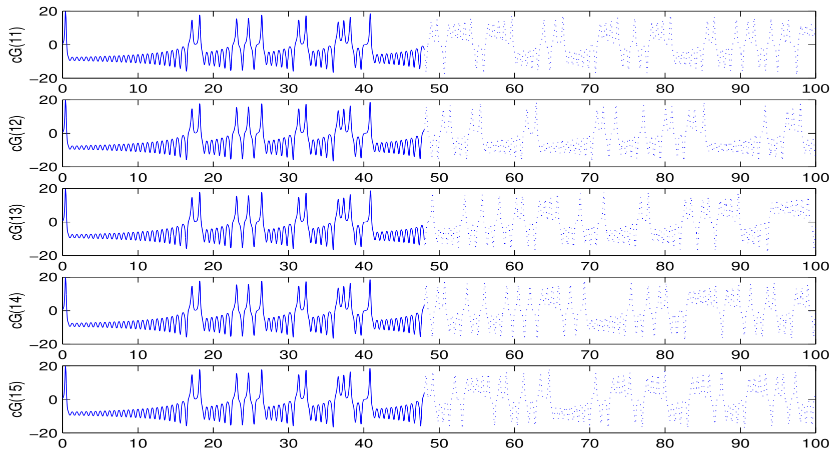


FIGURE 58.6. The U_1-component of cG(q) solutions for $q = 11, 12, 13, 14, 15$ with time step $k = 0.1$. Dashed lines indicate where the solution is no longer accurate.

Chapter 58 Problems

58.1. Verify that the three equilibrium points given in the text satisfy $f(u) = 0$. Linearise the system around these equilibrium points, i.e. compute the eigenvalues (and eigenvectors) for the Jacobian of f at the three equilibrium points.

58.2. Compute a solution to the Lorenz system and plot the orbit $(x(t), y(t), z(t)) = u(t)$ for $t \in [0, T]$. Do you agree with the description of the dynamics of the Lorenz system as never-ending process of cutting, expansion, flipping, and interlacing?

58.3. Repeat the experiment outlined in Section 58.4, i.e. compute solutions to the Lorenz system using the cG(1) method with a sequence of smaller and smaller time steps and examine the accuracy of the solutions (by comparing them to each other). Can you reach beyond $T = 25$?

58.4. Try the same experiment as in the previous problem but now with the lower order methods explicit Euler and implicit Euler. How far do you reach now?

58.5. Implement a simple version of the fourth-order cG(2) method given by

$$\begin{aligned} U(t_{n-1/2}) &= U(t_{n-1}) + \int_{t_{n-1}}^{t^n} f(U(t), t) \cdot (5 - 6(t - t_{n-1})/k_n)/4 \, dt, \\ U(t_n) &= U(t_{n-1}) + \int_{t_{n-1}}^{t^n} f(U(t), t) \, dt, \end{aligned}$$

where $U(t)$ is the quadratic polynomial on $[t_{n-1}, t_n]$ determined by the three values $U(t_{n-1})$, $U(t_{n-1/2})$ and $U(t_n)$. How much further can you reach with this method?

58.6. Give a motivation for the additional term in the refined error estimate (58.3), starting from the estimate containing errors caused by using the wrong initial condition as presented in Chapter *Adaptive IVP-solvers*.

58.7. Take on the Lorenz Challenge, i.e. compute an accurate solution over $[0, T]$ with T as large as possible. No rules, all is allowed!

59
The Solar System*

There is talk of a new astrologer who wants to prove that the earth moves and goes around instead of the sky, the sun, the moon, just as if somebody were moving in a carriage or ship might hold that he was sitting still and at rest while the earth and the trees walked and moved. But that is how things are nowadays: when a man wishes to be clever he must needs invent something special, and the way he does it must needs be the best! The fool wants to turn the whole art of astronomy upside-down. However, as Holy Scripture tells us, so did Joshua bid the sun to stand still and not the earth.
(Sixteenth century reformist M. Luther in his table book *Tischreden*, in response to Copernicus' pamphlet *Commentariolus*, 1514.)

59.1 Introduction

The problem of mathematical modeling of our Solar System including the Sun, the 9 planets Venus, Mercury, Tellus (the Earth), Mars, Jupiter, Saturn, Uranus, Neptune and Pluto together with a large number of moons and asteroids and occasional comets, has been of prime concern for humanity since the dawn of culture. The ultimate challenge concerns mathematical modeling of the Universe consisting of billions of galaxies each one consisting of billions of stars, one of them being our own Sun situated in the outskirts of the Milky Way galaxy.

According to the *geocentric* view presented by Aristotle (384-322 BC) in *The Heavens* and further developed by Ptolemy (87-150 AD) in *The Great*

System dominating the scene over 1800 years, the Earth is the center of the Universe with the Sun, the Moon, the other planets and the stars moving around the Earth in a complex pattern of circles upon circles (so-called epicycles). Copernicus (1473–1543) changed the view in *De Revolutionibus* and placed the Sun in the center in a new *heliocentric* theory, but kept the complex system of epicycles (now enlarged to a very complex system of 80 circles upon circles). Johannes Kepler (1572–1630) discovered, based on the extensive accurate observations made by the Swedish/Danish scientist Tycho Brahe (1546–1601), that the planets move in elliptic orbits with the Sun in one of the foci following *Kepler's laws*, which represented an enormous simplification and scientific rationalization as compared to the system of epicycles.

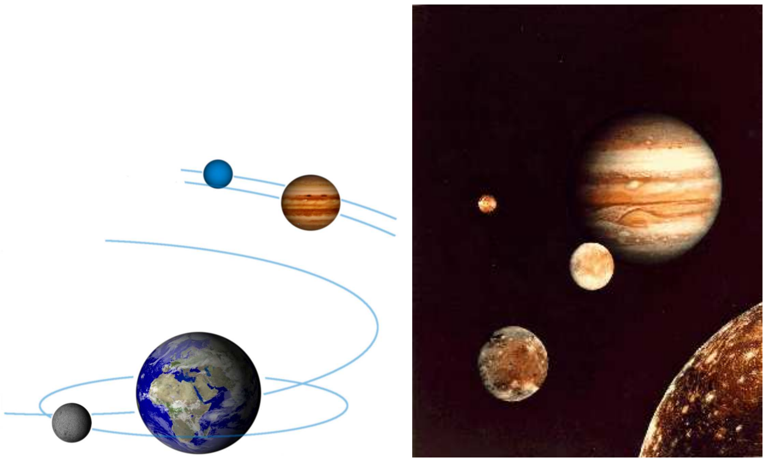


FIGURE 59.1. The Earth with the Moon and some other planets in orbit. Jupiter and the Galilean satellites, Io, Europa, Ganymede, and Callisto

In fact, already Aristarchus (310-230 BC) of Samos understood that the Earth rotates around its axis and thus could explain the (apparent) motion of the stars, but these views were rejected by Aristotle arguing as follows: if the Earth is rotating, how is it that an object thrown upwards falls on the same place? How come this rotation does not generate a very strong wind? No one until Copernicus could question these arguments. Can you?

Newton (1642–1727) then cleaned up the theory by showing that the motion of the planets could be explained from one single hypothesis: the inverse square law of gravitation, see Chapter *Newton's nightmare* below. In particular, Newton derived Kepler's laws for the *two-body problem* with one (small) planet in an elliptic orbit around a (large) sun, see Chapter *Lagrange and the Principle of Least Action.* Leibniz critized Newton for not giving any explanation of the inverse square law, which Leibniz believed could be derived from some basic fact, beyond one of "mutual love" which was quite popular. A sort of explanation was given by Einstein (1879–1955) in his theory of *General Relativity* with gravition arsing as a consequence of space-time being "curved" by the presence of mass. Einstein revolutionized *cosmology*, the theory of the Universe, but relativistic effects only add small corrections to Newton's model for our Solar System based on the inverse square law. Einstein gave no explanation why space-time gets curved by mass, and still today there is no convincing theory of gravitation with its mystical feature of "action at a distance" through some mechanism yet to be discovered. In Chapter *Laplacian models* below we give a derivation of the inverse square law using a mathematical argument presented by Laplace.

Despite the lack of a physical explanation of the inverse square law, Newton's theory gave an enourmous boost to mathematical sciences and a corresponding kick to the egos of scientists: if the human mind was capable of (so easily and definitely) understanding the secrets of the Solar System, then there could be no limits to the possibilities of scientific progress...

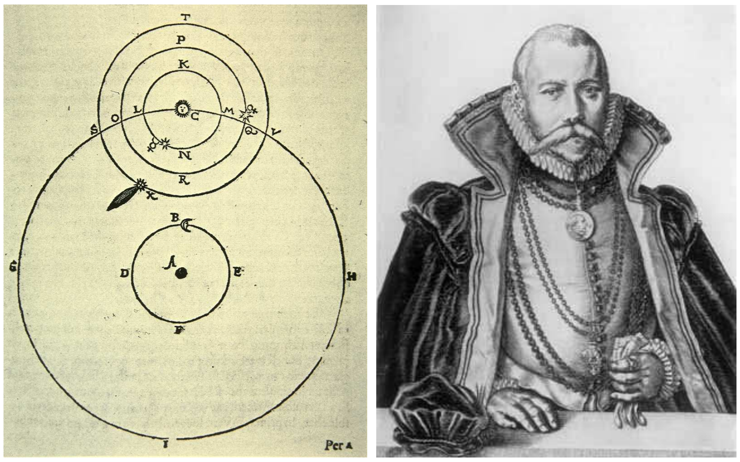


FIGURE 59.2. Tycho Brahe: "I believe that the Sun and the Moon orbit around the Earth but that the other planets orbit around the Sun".

59.2 Newton's Equation

The basis of celestial mechanics is Newton's second law,

$$F = m \cdot a, \tag{59.1}$$

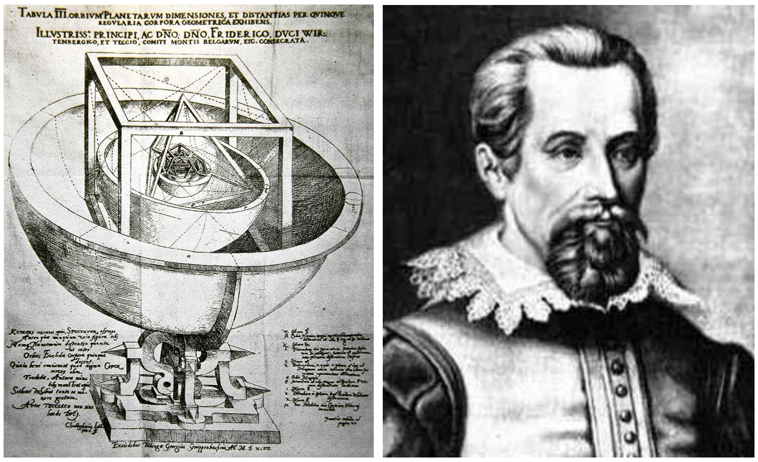


FIGURE 59.3. Johannes Kepler: "I believe that the planets are separated by invisible regular polyhedra: tetrahedron, cube, octahedron, dodekahedron and ikosahedron, and further that the planets including the Earth move in elliptical orbits around the Sun".

expressing that a *force* F results in an acceleration of size a for a body of mass m, together with the expression for the gravitational force given by the inverse square law:

$$F = G\frac{mm_a}{r^2}, \tag{59.2}$$

where $G \approx 6.67 \cdot 10^{-11}\mathrm{Nm}^2/\mathrm{kg}^2$ is the *gravitational constant*, m_a is the mass of the attracting body and r is the distance to the attracting body.

Together (59.1) and (59.2) give a set of differential equations for the evolution of the Solar System. If we know the initial positions and velocities for all bodies in the Solar System, we can solve the system of differential equations, using the same techniques as presented above in Chapter Adaptive IVP-Solvers. We discuss this in detail below in Section 59.4. As a preparation, we rewrite (59.1) and (59.2) in dimensionless form, which will be convenient. The three fundamental units appearing in the equations are those of *space*, *time* and *mass*, which are represented by the variables x (or r), t and m. We now introduce new dimensionless variables, $x' = x/\mathrm{AU}$, $t' = t/\mathrm{year}$ and $m' = m/M$, where 1 AU is the mean distance from the Sun to Earth and M is the mass of the Sun. We can use the chain rule to obtain the dimensionless acceleration, $a' = \frac{d}{dt'}\frac{d}{dt'}x' = \frac{dt}{dt'}\frac{d}{dt}\frac{dt}{dt'}\frac{d}{dt}x/\mathrm{AU} = \frac{\mathrm{year}^2}{\mathrm{AU}}a$. Combining (59.1) and (59.2) using our new dimensionless variables, we then obtain

$$m'M\frac{\mathrm{AU}}{\mathrm{year}^2}a' = G\frac{m'M \cdot m'_a M}{r'^2\mathrm{AU}^2}, \tag{59.3}$$

or

$$a' = G' \frac{m'_a}{r'^2}, \tag{59.4}$$

where the new gravitational constant G' is given by

$$G' = \frac{G \cdot \text{year}^2 M}{\text{AU}^3}. \tag{59.5}$$

We leave it as an exercise to show that with suitable definitions of the units year and AU, the new dimensionless gravitational constant G' is given by

$$G' = 4\pi^2. \tag{59.6}$$

59.3 Einstein's Equation

In general relativity the basic concept is not *force*, as in Newtonian theory, but instead the *curvature* of space-time. Einstein explains the motion of the planets in our Solar System in the following way: the planets move through space-time along straight lines, *geodesics*, which appear as circular (or elliptical) orbits only because space-time is curved by the large mass of the Sun. We shall now try to give an idea of how this works.

The curvature of space-time is given by its *metric*. A metric defines the distance between two nearby points in space-time. In Euclidean geometry that we have studied extensively in this book, the distance between to points $x = (x_1, x_2, x_3)$ and $y = (y_1, y_2, y_3)$ is given by the square root of the scalar product $dx \cdot dx$, where dx is the difference $dx = x - y$. With the notation $ds = |x - y|$ we thus have

$$ds = \sqrt{dx \cdot dx} = \left(\sum_{i=1}^{3} dx_i^2 \right)^{1/2}, \tag{59.7}$$

or

$$ds^2 = \sum_{i=1}^{3} dx_i^2. \tag{59.8}$$

In the notation of general relativity, the Euclidean metric is then given by the matrix (tensor)

$$g = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{59.9}$$

as

$$ds^2 = dx^T g\, dx. \tag{59.10}$$

In space-time we include time t as a fourth coordinate and every event in space-time is given by a vector (t, x_1, x_2, x_3). In flat or *Minkowski* space-time in the absence of masses, the curvature is zero and the metric is given by

$$g = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{59.11}$$

which gives

$$ds^2 = -dt^2 + dx_1^2 + dx_2^2 + dx_3^2. \tag{59.12}$$

In the precense of masses, we obtain a different metric which does not even have to be diagonal.

From the metric g one can find the straight lines of space-time, which give the orbits of the planets. The metric itself is determined by the distribution of mass in space-time, and is given by the solution of Einstein's equation,

$$R_{ij} - \frac{1}{2} R g_{ij} = 8\pi T_{ij}, \tag{59.13}$$

where (R_{ij}) is the so-called *Ricci-tensor*, R is the so-called *scalar curvature* and (T_{ij}) is the so-called *stress-energy tensor*. Now (R_{ij}) and R depend on derivatives of the metric $g = (g_{ij})$ so (59.13) is a partial differential equation for the metric g.

The solution for the orbits of the planets obtained from Einstein's equation are a little different than the solution obtained from (59.4) given by Newton. Although the difference is small, it has been verified in observations of the orbit of the planet Mercury which is the planet closest to the Sun. We will not include these "relativistic effects" in the next section where we move on to the computation of the evolution of the Solar System.

59.4 The Solar System as a system of ODEs

To use the techniques developed in Chapter Adaptive IVP-Solvers to compute the evolution of the Solar System, we need to rewrite the second-order system of ODEs given by (59.4) in the standard form $\dot{u} = f$. We start by introducing coordinates $x^i(t) = (x_1^i(t), x_2^i(t), x_3^i(t))$ for all bodies in the Solar System, including the nine planets, then Sun and the Moon. This gives a total of $n = 9 + 2 = 11$ bodies and a total of $3n = 33$ coordinates. To rewrite the equations as the first-order system $\dot{u} = f$ we need to include also the velocities of all bodies, $\dot{x}^i(t) = (\dot{x}_1^i(t), \dot{x}_2^i(t), \dot{x}_3^i(t))$, giving a total of $N = 6n = 66$ coordinates. We collect all these coordinates in the vector $u(t)$ of length N in the following order:

$$\begin{aligned} u(t) = (&x_1^1(t), x_2^1(t), x_3^1(t), \ldots, x_1^n(t), x_2^n(t), x_3^n(t), \\ &\dot{x}_1^1(t), \dot{x}_2^1(t), \dot{x}_3^1(t), \ldots, \dot{x}_1^n(t), \dot{x}_2^n(t), \dot{x}_3^n(t)), \end{aligned} \tag{59.14}$$

so that the first half of the vector $u(t)$ contains the positions of all bodies and the second half contains the corresponding velocities.

To obtain the differential equation for $u(t)$, we take the time-derivative and notice that the derivative of the first half of $u(t)$ is equal to the second half of $u(t)$:

$$\dot{u}_i(t) = u_{3n+i}(t), \quad i = 1, \ldots, 3n, \tag{59.15}$$

i.e. for $n = 11$ we have $\dot{u}_1(t) = \dot{x}_1^1(t) = u_{34}(t)$ and so on.

The derivative of the second half of $u(t)$ will contain the second derivatives of the positions, i.e. the accelerations, and these are given by (59.4). Now (59.4) is written as a scalar equation and we have to rewrite it in vector form. For every body in the Solar System, we need to compute the contribution to the total force on the body by summing the contributions from all other bodies. Assuming that we work in dimensionless variables (but writing x instead of x', m_i instead of m_i' and so on for convenience) we then need to compute the sum:

$$\ddot{x}_i(t) = \sum_{j \neq i} \frac{G' m_j}{|x^j - x^i|^2} \frac{x^j - x^i}{|x^j - x^i|}, \tag{59.16}$$

where the unit vector $\frac{x^j - x^i}{|x^j - x^i|}$ gives the direction of the force, see Figure 59.4.

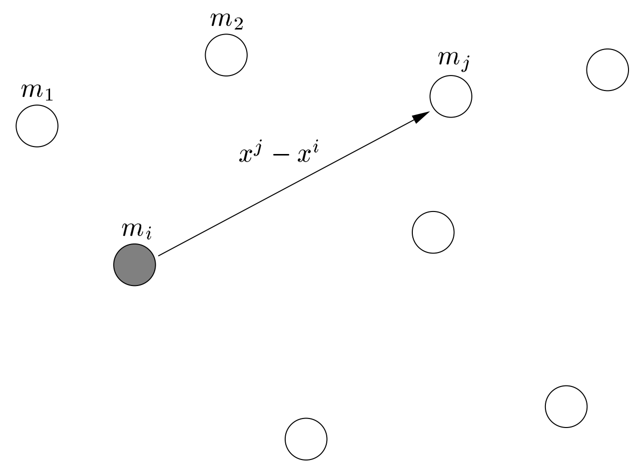


FIGURE 59.4. The total force on body i is the sum of the contributions from all other bodies.

Our final differential equation for the evolution of the Solar System in the form $\dot{u} = f$ is then given by

$$\dot{u}(t) = f(u(t)) = \begin{bmatrix} u_{3n+1}(t) \\ \vdots \\ u_{6n}(t) \\ \sum_{j \neq 1} \frac{G' m_j}{|x^j - x^1|^2} \frac{x_1^j - x_1^1}{|x^j - x^i|} \\ \vdots \\ \sum_{j \neq n} \frac{G' m_j}{|x^j - x^n|^2} \frac{x_3^j - x_3^n}{|x^j - x^n|} \end{bmatrix}, \qquad (59.17)$$

where we have kept the notation $x^1 = (x_1^1, x_2^1, x_3^1)$ rather than (u_1, u_2, u_3) and so on in the right-hand side for simplicity. The evolution of our Solar System can now be computed by the standard techniques developed in Chapter Adaptive IVP-Solvers, using the initial data supplied in Table 59.1.

59.5 Predictability and Computability

Two important questions that arise naturally when we study numerical solutions of the evolution of our Solar System, such as the one in Figure 59.5, are the questions of *predictability* and *computability*.

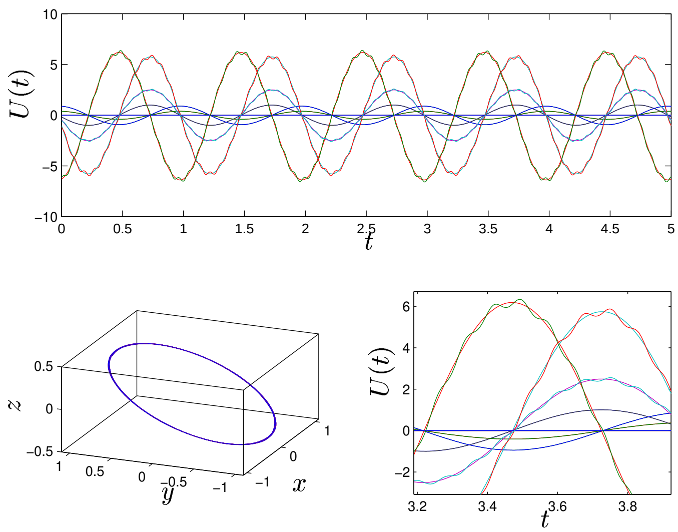


FIGURE 59.5. A numerical computation of the evolution of the Solar System, including Earth, the Sun and the Moon.

	Position		Velocity		Mass
90Mercury	$x^1(0) =$	-0.147853935 -0.400627944 -0.198916163	$\dot{x}^1(0) =$	7.733816715 -2.014137426 -1.877564183	1.0/6023600
90Venus	$x^2(0) =$	-0.725771746 -0.039677000 0.027897127	$\dot{x}^2(0) =$	0.189682646 -6.762413869 -3.054194695	1.0/408523.5
90Earth	$x^3(0) =$	-0.175679599 0.886201933 0.384435698	$\dot{x}^3(0) =$	-6.292645274 -1.010423954 -0.438086386	1.0/328900.5
90Mars	$x^4(0) =$	1.383219717 -0.008134314 -0.041033184	$\dot{x}^4(0) =$	0.275092348 5.042903370 2.305658434	1.0/3098710
90Jupiter	$x^5(0) =$	3.996313003 2.731004338 1.073280866	$\dot{x}^5(0) =$	-1.664796930 2.146870503 0.960782651	1.0/1047.355
90Saturn	$x^6(0) =$	6.401404019 6.170259699 2.273032684	$\dot{x}^6(0) =$	-1.565320566 1.286649577 0.598747577	1.0/3498.5
90Uranus	$x^7(0) =$	14.423408013 -12.510136707 -5.683124574	$\dot{x}^7(0) =$	0.980209400 0.896663122 0.378850106	1.0/22869
90Neptune	$x^8(0) =$	16.803677095 -22.983473914 -9.825609566	$\dot{x}^8(0) =$	0.944045755 0.606863295 0.224889959	1.0/19314
90Pluto	$x^9(0) =$	-9.884656563 -27.981265594 -5.753969974	$\dot{x}^9(0) =$	1.108139341 -0.414389073 -0.463196118	1.0/150000000
90Sun	$x^{10}(0) =$	-0.007141917 -0.002638933 -0.000919462	$\dot{x}^{10}(0) =$	0.001962209 -0.002469700 -0.001108260	1
90Moon	$x^{11}(0) =$	-0.177802714 0.884620944 0.384016593	$\dot{x}^{11}(0) =$	-6.164023246 -1.164502534 -0.506131880	$1.0/2.674 \cdot 10^7$

TABLE 59.1. Initial data for the Solar System at 00.00 Universal Time (UT1, approximately GMT) January 1 2000 for dimensionless positions and velocities scaled with units 1 AU $= 1.49597870 \cdot 10^{11}$ m (one astronomical unit), 1 year $= 365.24$ days and $M = 1.989 \cdot 10^{30}$ kg (one solar mass).

The predictability of the Solar System is the question of the accuracy of a computation given the accuracy in initial data. If initial data is known with an accuracy of say five digits, and the numerical computation is exact, how long does it take until the solution is no longer accurate even to one digit?

The computability of the Solar System is the question of the accuracy in a numerical solution given exact initial data, i.e. how far we can compute an accurate solution with available resources such as method, computational power and time.

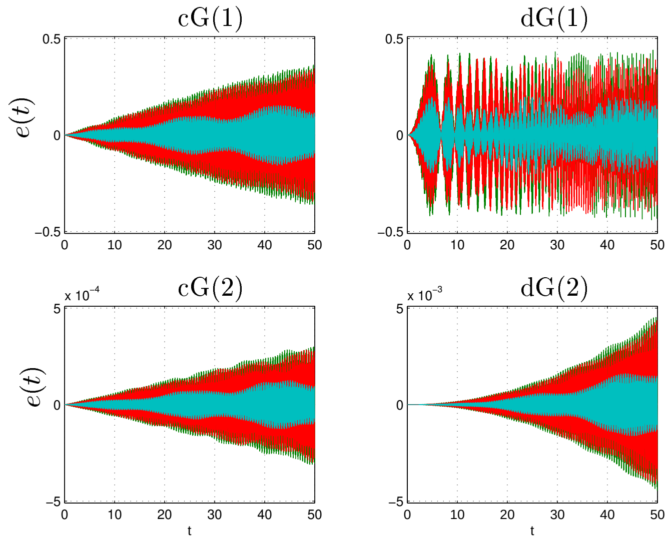


FIGURE 59.6. The growth of the numerical error in simulations of the Solar System using different numerical methods. The two methods on the left conserve energy, which results in linear rather than quadratic error growth.

Both the predictability and the computability are determined by the growth rate of errors. Luckily, the error does not grow exponentially as we saw for the Lorenz system. If we imagine that we displace Earth slightly from its orbit and start a computation, the orbit and velocity of Earth will be slightly different, resulting in an error that grows *linearly* with time. This means that the predictability of the Solar System is quite good, since every extra digit of accuracy in initial data means that the limit of predictability is increased by a factor ten. If now the solution is computed using a numerical method, such as the adaptive cG(1) method, this will result in additional errors. We can think of the error caused by a numerical method as a small perturbation introduced with every new time step. Adding the contributions from all time steps we find that the numerical error grows *quadratically*, see Problem 59.2.

As it turns out however, the error does not grow quadratically but only linearly for the cG(1) method as shown in Figure 59.6. This pleasant surprise is the result of an important property of the cG(1) method: it conserves energy. As a result, the cG(1) method performs better on a long time interval than the higher-order (more accurate) dG(2) method.

59.6 Adaptive time-stepping

If we compute the evolution of the Solar System using the adaptive cG(1) method, we find that the time steps need to be small enough to follow the orbit of the Moon (or Mercury if we do not include the Moon). This is inefficient since the time scales for the other bodies are much larger: the period of the Moon is one month and the period of Pluto is 250 years, and so the time steps for Pluto should be roughly a factor 3,000 larger that the time steps for the Moon. It has been shown recently that the standard methods cG(q), including cG(1), and dG(q) can be extended to individual, *multi-adaptive*, time-stepping for different components. In Figure 59.7 we show a computation made with individual time steps for the different planets. Notice how the error grows quadratically, indicating that the method does not conserve energy. (It is possible to construct also multi-adaptive methods which conserve energy.)

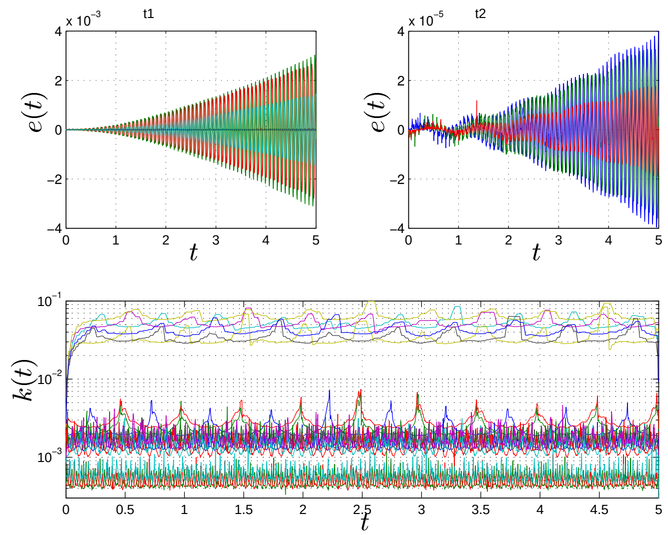


FIGURE 59.7. A computation of the evolution of the Solar System with individual, multi-adaptive, time steps for the different planets.

59.7 Limits of computability and predictability

Using the multiadaptive cG(2) it appears that the limit of computability of the Solar System (with the Moon and the nine planets) using double precision, is of the order 10^6 years. Concerning the predictability of the same system it appears that for every digit beyond 5 in the precision of data we gain a factor of ten in time, so that for example predicting the position of the Moon 1000 years ahead would require about 8 correct digits in e.g. the initial positions and velocities, masses and gravitational constant. We conclude that it appears that normally the precision in data would set the limit for accurate simulations of the evolution of the Solar System, if we use a high order multiadaptive solver.

Chapter 59 Problems

59.1. Prove that with suitable definitions of the units year and AU the gravitational constant is $G' = 4\pi^2$. *Hint:* assume that Earth is in a circular orbit where the centripetal force mv^2/r is balanced by the gravitational force GmM/r^2.

59.2. Motivate the quadratic growth of the numerical error for the Solar System. *Hint:* Assume that an error of size ϵ is added to the velocity of a planet in every time step.

59.3. (*Hard!*) Prove that in general if an error in initial data grows as

$$|e(T)| \leq S(T)|e(0)|,$$

for a specific initial value problem, then the error in a numerical solution of the initial value problem grows as

$$|e(T)| \lessapprox \epsilon \int_0^T S(t)\, dt,$$

assuming that the additional error in every time step is kept below $k_n\epsilon$.

59.4. Prove that the cG(1) method conserves energy for a Hamiltonian system, i.e. prove that for a system given by $\ddot{x} = F = -\nabla_x P(x)$, the total energy

$$E(t) = K(\dot{x}(t)) + P(x(t)),$$

is conserved. Here $P(x)$ is a given potential field, and $K(\dot{x}) = \frac{\dot{x}^2}{2}$ is the kinetic energy. *Hint:* Write as a first-order system for the vector $[u, v] = [x, \dot{x}]$, take $[\dot{v}, \dot{u}]$ as the test function and use the chain rule.

59.5. Investigate numerically the predictability and computability of the Solar System. Can you verify the linear error growth for the cG(1) method?

59.6. Add a comet to your Solar System and try to choose the initial data for the comet so that it (almost) hits the Moon. Can you make the Moon come loose from Earth? Do errors still grow linearly?

> But Aristarchus has brought out a book consisting of certain hypotheses, wherein it appears, as a consequence of the assumptions made, that the universe is many times greater than the 'universe' just mentioned. His hypotheses are that the fixed stars and the sun remain unmoved, that the earth revolves about the sun on the circumference of a circle, the sun lying in the middle of the orbit, and that the sphere of fixed stars, situated about the same centre as the sun, is so great that the circle in which he supposes the earth to revolve bears such a proportion to the distance of the fixed stars as the centre of the sphere bears to its surface. (Archimedes about Aristharcus of Samos)

60 Optimization

1. All living beings are driven by passion to seek maximal Pleasure.
2. There is Pleasure of the Body and Pleasure of the Soul. In the Pleasure of the Soul, the Body cannot take part, whereas the Pleasure of the Body is equally shared by the Soul. (the 2 first of the 14 basic principles of *Anthropologica physica*, by King Karl XII of Sweden, 1717)

60.1 Introduction

In this chapter, we expand on some basic aspects of optimization touched upon in the previous chapter in connection to minimization. Optimization is very rich subject and we shall return to other aspects below. The issues we discuss here are connected to the very basics of Calculus and are considered as "deep" and understandable only by the very best math majors. You may test your own reaction to the discussions presented, and if you get the expected feeling of confusion, don't worry, just proceed to the next chapter. If on the other hand, against all odds, you get the feeling of grasping the main ideas, then you may congratulate yourself for being more gifted for mathematics than you thought!

In our modern world, *optimization* is a code word. To *optimize* is to use available resources as efficiently as possible, or to find the best of available alternatives. In our private lives, we may want our car to use as little fuel as possible, to buy an item at lowest possible price, to use as little effort as possible to clean the house, or to get maximal enjoyment out of the vacation trip.

In automatized production, the leading principle is always to optimize and seek to use as little energy, material and human resources as possible to produce a certain amount of goods. A basic idea in our capitalistic system is that in the long run the most efficient mode of production will win the market.

A basic problem of optimization is to find the maximum or minimum value of a given function $f : \Omega \to \mathbb{R}$ defined on some set of numbers Ω. Typically, Ω may be a domain in $\mathbb{R}^d$ with $d = 1, 2, 3, ...$, that may be bounded or unbounded, or Ω may be a finite set such as the set of natural numbers $1, 2..., 100$. More precisely, finding a *minimum point* $\bar{x}$ in Ω amounts to finding a point $\bar{x} \in \Omega$ such that

$$f(x) \geq f(\bar{x}) \quad \text{for all } x \in \Omega, \tag{60.1}$$

and we then say that $f(\bar{x})$ is the *minimal value* of $f : \Omega \to \mathbb{R}$. Note that there may be several minimum points, but of course there may be only one minimal value. If in an Olympic 100 meter race, three runners share the best time of 9.99 seconds, then all the three runners may share the gold medal. However, there cannot be two runners with different final times who both get a gold medal.

We now consider the problem of finding the minimum value and corresponding minimum point(s) of a given function $f : \Omega \to \mathbb{R}$. We may distinguish between the following two cases: (a) Ω is a domain of $\mathbb{R}^d$ with infinitely many points, as when Ω is the unit disc $\{x \in \mathbb{R}^d : \|x\| \leq 1\}$; (b) Ω contains finitely many points, as for example $\Omega = \{1, 2, 3, ..., 10\}$. The case (a) is "continuous" and (b) is "discrete". The two cases are not fully disjoint; there may be a gradual passage from "discrete" to "continuous" as the number of elements in Ω increases. In the case Ω is discrete with finitely many points, we may find the minimum value and corresponding minimum point(s) of $f : \Omega \to \mathbb{R}$ by different algorithms for *sorting.* If Ω is "continuous" with infinitely many points, sorting may be impossible and different algorithms that use information from the derivative of $f(x)$ in variations of steepest descent are often used.

60.2 Sorting if Ω is finite

If Ω is a finite set of numbers, for example if $\Omega = \{1, 2, ..., 9, 10\}$, then we just make a list of the corresponding 10 function values $f(1)$, $f(2)$,...,$f(10)$ and by sorting these values according to magnitude in increasing order, we can find the minimal value $f(\bar{x})$ and the corresponding argument $\bar{x}$. Of course, we don't have to sort all the numbers according to magnitude to find the smallest one. We just have to find the first element in the list sorted according to magnitude in increasing order. Repeating this process we can sort all the numbers in the given list of numbers.

EXAMPLE 60.1. For example, suppose that $\Omega = \{1, 2, ..., 9, 10\}$ and $f(1) = 143$, $f(2) = 538$, $f(3) = 67$, $f(4) = 964$, $f(5) = 287$, $f(6) = 64$, $f(7) = 123$, $f(8) = 333$, $f(9) = 63$, $f(10) = 88$. By direct inspection, we see that the minimum point is $\bar{x} = 9$ and the minimum value is $f(\bar{x}) = 63$.

While sorting sounds simple, it turns out to be an interesting problem to do sorting efficiently when there are a large number of values to be sorted. So there are different algorithms for sorting and sorting algorithms hold a prominent place in computer science. A simple algorithm for finding the minimum m of N numbers $f(1), \cdots, f(N)$ goes as follows:

1. Set $m = f(1)$ and $\bar{x} = 1$
2. For $x = 2, \cdots, N$, if $f(x) < m$ set $m = f(\bar{x})$ and $\bar{x} = x$.

The minimum value is then $m = f(\bar{x})$ and the minimum point is $x = \bar{x}$. The algorithm is based on successive comparison of pairs of numbers (if $f(x) < m$ then we update m and $\bar{x}$ and set $m = f(\bar{x})$ and $\bar{x} = x$). The number of comparisons in the indicated algorithm is apparently $N - 1$.

Repeating the algorithm with $f(\bar{x})$ eliminated, we can get a complete sorting according to magnitude using $(N-1) + (N-2) + \cdots + 1 \approx \frac{1}{2}N^2$ comparisons.

60.3 What if Ω is not finite?

If Ω is an interval of real numbers, for example $\Omega = [0, 1]$, then Ω contains infinitely many points and sorting the values $f(x)$ with $x \in \Omega$ of a given function $f : \Omega \to \mathbb{R}$ by pairwise comparison appears impossible in practice because we cannot perform infinitely many comparisons. Of course, in practice we replace Ω by a finite set of numbers, for example by using a single precision floating point representation of the points in Ω. So in principle, we can then apply the above sorting strategy. But, the procedure will be computationally intensive. With seven digits we would have 10^7 values $f(x)$ to compare, which using the above algorithm requires on the order of 10^7 comparisons to find the minimum. If the interval Ω is larger and the desired precision higher, then the number of comparisons would be correspondingly larger. The total computational cost would also involve as a multiplicative factor the cost of evaluating the function value $f(x)$ for a given x, which itself could require many arithmetic operations. The total cost in direct comparison thus may be prohibitively large.

We now seek efficient algorithms to handle the case that Ω is a domain of $\mathbb{R}^d$ interval and the function $f(x)$ is Lipschitz continuous with a Lipschitz constant L. In this case, the function values $f(x)$ cannot change more than the argument of x changes times L. If we want to find the minimal value up to a certain tolerance TOL, then we need to do approximately $(L/TOL)^d$ comparisons if the diameter of Ω is of order one. Depending on the choice of the tolerance TOL and L this may be acceptable or not.

If the function $f(x)$ is differentiable then we may restrict the search further using information from the derivative, as we shall see below.

60.4 Existence of minimum point

How can we be sure that there in fact is a minimum point? We discuss the proof of the following basic theorem addressing this question below.

Theorem 60.1 *If $f : \Omega \to \mathbb{R}$ is Lipschitz continuous, where Ω is a closed and bounded subset of $\mathbb{R}^d$, then there is a minimum point $\bar{x} \in \Omega$ such that $f(\bar{x}) \leq f(x)$ for all $x \in \Omega$.*

The assumption that Ω is closed and bounded is essential to guarantee existence of the minimum, as the following examples show.

EXAMPLE 60.2. The function $f : (0, 1) \to \mathbb{R}$ with $f(x) = x$ does not have a minimum point in $(0, 1)$. In this case $\Omega = (0, 1)$ is not closed.

EXAMPLE 60.3. The function $f : [1, \infty) \to \mathbb{R}$ with $f(x) = 1/x$ does not have a minimum point in $[1, \infty)$. In this case $\Omega = [1, \infty)$ is not bounded.

Note however that a function $f : \Omega \to \mathbb{R}$ may have a minimum even if Ω is unbounded. In particular, if $f(x)$ increases to infinity as $\|x\|$ increases, then we can effectively reduce the search for a minimum to a bounded set.

EXAMPLE 60.4. The function $f : [0, \infty)$ given by $f(x) = x^2 - 2x$ attains a minimum value $f(1) = -1$; since $f(x) \geq 0$ for $x \geq 2$, we may restrict the search for a minimum to $[0, 2]$.

60.5 The derivative is zero at an interior minimum point

We assume that $f : \Omega \to \mathbb{R}$ is a given Lipschitz continuous differentiable function, where Ω is a domain in $\mathbb{R}^d$. We shall now prove that if $\bar{x}$ is an *interior minimum point* of $f : \Omega \to \mathbb{R}$, that is $\bar{x}$ is a minimum point and the ball $\{x \in \mathbb{R}^d : \|x - \bar{x}\| < \delta\}$ is included in Ω for some positive number δ, then $f'(\bar{x}) = \nabla f(\bar{x}) = 0$, where $f' = \nabla f$ is the gradient of f. This follows by writing

$$f(x) = f(\bar{x}) + f'(\bar{x}) \cdot (x - \bar{x}) + E_f(x, \bar{x})$$

with $|E_f(x, \bar{x})| \leq K_f(\bar{x})\|x - \bar{x}\|^2$. If now $f'(\bar{x}) \neq 0$, we may choose $x = \bar{x} - \epsilon f'(\bar{x}) \in \Omega$ with $\epsilon > 0$ and estimate to get

$$\begin{aligned} f(x) &\leq f(\bar{x}) - \epsilon\|f'(\bar{x})\|^2 + \epsilon^2 K_f(\bar{x})\|f'(\bar{x})\|^2 \\ &= f(\bar{x}) - \epsilon\|f'(\bar{x})\|^2(1 - \epsilon K_f(\bar{x})) < f(\bar{x}). \end{aligned}$$

For ϵ sufficiently small, we get a contradiction to the assumption that $\bar{x}$ is a minimum point. We have proved the following basic result, see Fig. 60.1 and Fig. 60.2.

Theorem 60.2 *Suppose $f : \Omega \to \mathbb{R}$ has a minimum point at an interior point $\bar{x}$ in Ω, and suppose that $f : \Omega \to \mathbb{R}$ is differentiable at $\bar{x}$. Then* $f'(\bar{x}) = 0$.

Using this result, we may search for interior minimum points among the zeros of the derivative $f'(x)$ in Ω. To find these zeros we may use some algorithm for computing roots, like Fixed Point Iteration or Newton or the Bisection algorithm. There is thus a strong connection between algorithms for finding interior minimum points of $f : \Omega \to \mathbb{R}$ and algorithms for computing roots of $f'(x) = 0$.

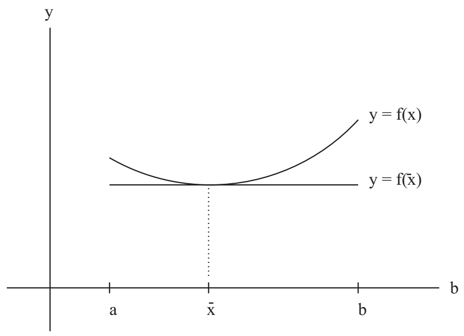


FIGURE 60.1. $f'(\bar{x}) = 0$ at an interior minimum point $\bar{x}$.

Note that if the minimum point $\bar{x}$ of $f : \Omega \to \mathbb{R}$ is not interior to Ω, i.e. $\bar{x}$ lies on the boundary of Ω, then the derivative $f'(\bar{x})$ may be non-zero, see Fig. 60.3.

EXAMPLE 60.5. Suppose we want to minimize $f : \Omega \to \mathbb{R}$ with $\Omega = [0, 2]$ and $f(x) = x^2 - 2x$. Since Ω is closed and bounded and $f(x)$ is Lipschitz continuous, we know that there is a minimum point $\bar{x} \in [0, 2]$. If $\bar{x}$ is interior to $[0, 2]$, that is if $0 < \bar{x} < 2$, then $f'(\bar{x}) = 2\bar{x} - 2 = 0$ and thus $\bar{x} = 1$. We compare the value $f(1) = -1$ to the values $f(0) = 0$ and $f(2) = 0$ on the boundary of $[0, 2]$ and conclude that $f(1) = -1$ is the minimum value and $\bar{x} = 1$ the corresponding minimum point.

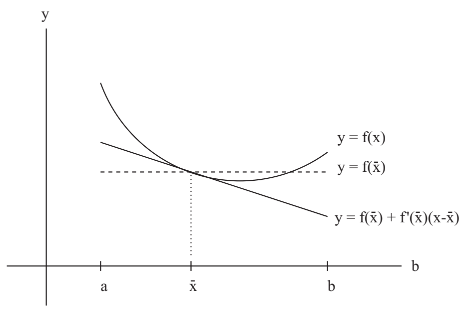


FIGURE 60.2. $f'(\bar{x}) < 0$ implies that $f(x) < f(\bar{x})$ for x close to $\bar{x}$ with $\bar{x} > x$, that is, $\bar{x}$ cannot be a minimum point.

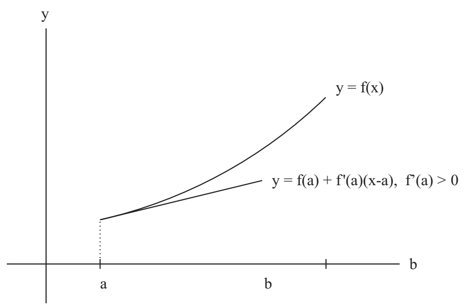


FIGURE 60.3. $f'(\bar{x})$ may be non-zero at a minimum $\bar{x}$ on the boundary.

EXAMPLE 60.6. Suppose we want to minimize $f : Q \to \mathbb{R}$ with $f(x) = f(x_1, x_2) = x_1^2 + x_2^2 - 2x_1 - x_2$ on a closed square $Q = [0,2] \times [0,2]$, see Fig. 60.4. We know that there is a minimum point in Q. We first compute the interior points $\hat{x}$ where $f'(\hat{x}) = 0$. Since $f'(x) = (2x_1 - 2, 2x_2 - 1)$, $\hat{x} = (1, 0.5)$ with the function value $f(1, 0.5) = -1.25$. It remains to study the variation of $f(x)$ on the boundary of Q to see if we find a value smaller than -1.25. We do this by considering each piece of the boundary separately. On the part $x_2 = 0$, we have $f(x) = x_1^2 - 2x_1$ with $x_1 \in [0,2]$, and we see arguing as in the previous example that the minimal value is $f(1,0) = -1$. On the part $x_2 = 2$, we have $f(x_1, 2) = x_1^2 - 2x_1 + 3$ with minimum $f(1,2) = 2$. On the part $x_1 = 0$, we have $f(0, x_2) = x_2^2 - x_2$ with minimum $f(0, 0.5) = -0.25$, and on the part $x_1 = 2$, we have $f(2, x_2) = x_2^2 - x_2$ with minimum $f(2, 0.5) = -0.25$. We conclude that the minimum point is the interior point $\bar{x} = (1, 0.5)$ and that the minimum value is $f(1, 0.5) = -1.25$.

EXAMPLE 60.7. You are asked to design a box (without top) of a given volume using as little material as possible. Letting the sides of the box

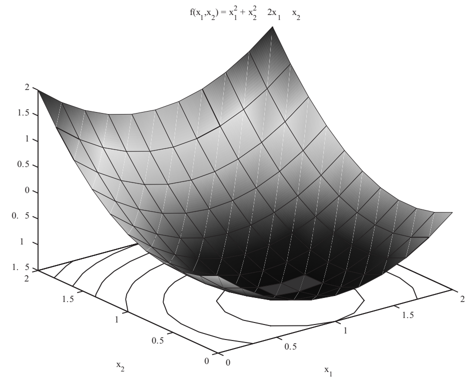


FIGURE 60.4. Minmizing $f(x) = x_1^2 + x_2^2 - 2x_1 - x_2$ on $Q = [0,2] \times [0,2]$

be x_1, x_2 and x_3, the volume is $x_1x_2x_3 = V$ and the surface to be minimized is $x_1x_2 + 2x_1x_3 + 2x_2x_3$. Eliminating x_3 gives

$$f(x_1, x_2) = x_1x_2 + 2V(\frac{1}{x_1} + \frac{1}{x_2}),$$

which is to be minimized over $\Omega = [0, \infty) \times [0, \infty)$. Seeking points $\hat{x}$ with $f'(\hat{x}) = (0,0)$, we find $\hat{x}_1 = \hat{x}_2 = (2V)^{1/3}$ with the corresponding height $\hat{x}_3 = \frac{1}{2}(2V)^{1/3}$, and the area

$$f(\hat{x}) = (2V)^{2/3} + 2(2V)^{2/3}.$$

Comparing with (x_1, x_2) with x_1 or x_2 very large or small give large values to $f(x_1, x_2)$ and thus the minimum point is $\hat{x}$. The solution is a box with square bottom and height half of the width.

We also remark that a minimum value may be attained at an interior point where the given function is *nondifferentiable*. For example, the minimum value of the function $f(x) = |x - 1|$ on $[0, 2]$ is attained at $\bar{x} = 1$ with minimum value $f(\bar{x}) = f(1) = 0$. This type of interior minimum points must be considered separately. Thus, to find all possible minimum points we have to consider the points $\bar{x}$ for which $f'(\bar{x}) = 0$, and in addition to these the end points of the domain of definition and interior points where $f(x)$ is nondifferentiable, see Fig. 60.5.

60.6 The role of the Hessian

We know that if $\bar{x}$ is an interior minimum point of a function $f : \Omega \to \mathbb{R}$, then $f'(\bar{x}) = 0$. But it is not true in general that if $f'(\bar{x}) = 0$, then $\bar{x}$ is a minimum point. A point $\bar{x}$ with $f'(\bar{x}) = 0$ may e.g. be a *maximum point*,

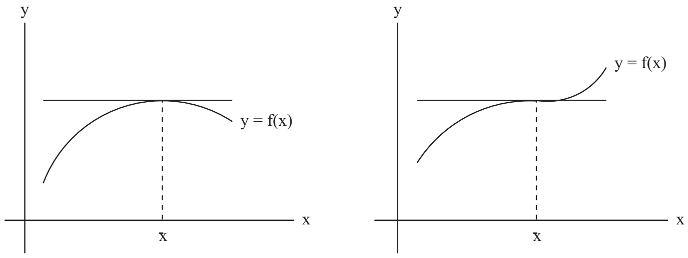


FIGURE 60.5. $f'(\bar{x}) = 0$ may correspond also to a maximum point or an inflection point.

or an *inflection point*, see Fig. 60.5. If the Hessian H of $f : \Omega \to$ is positive definite close to $\bar{x}$ and $f'(\bar{x}) = 0$, then we have by Taylor's theorem

$$f(x) = f(\bar{x}) + \frac{1}{2}(x - \bar{x})^\top H(y) \cdot (x - \bar{x}) > f(\bar{x})$$

for x close to $\bar{x}$ and spme y between x and $\bar{x}$, and thus $\bar{x}$ is a *local minimum point.*

We recall that an $n \times n$ matrix A is said to be positive definite if

$$v^\top Av > 0$$

for all non-zero $v \in \mathbb{R}^n$. The Spectral Theorem implies that A is positive definite if and only if the eigenvalues of A are positive.

EXAMPLE 60.8. If $A = (a_{ij})$ is a symmetric 2×2 matrix, then A is positive definite if

$$a_{11}a_{22} - a_{12}^2 > 0 \quad \text{and } a_{11} > 0.$$

This follows by completing squares in

$$v^\top Av = a_{11}v_1^2 + a_{22}v_2^2 + 2a_{12}v_1v_2.$$

60.7 Minimization algorithms: steepest descent

We discuss briefly how to find candidates for minimum points of a given function $f : \Omega \to \mathbb{R}$, where Ω is a domain in $\mathbb{R}^d$. We assume that $f : \Omega \to \mathbb{R}$ is Lipschitz continuous and differentiable on Ω.

In the *steepest descent method*, we construct a sequence $\{x_i\}$ in $\mathbb{R}^d$ that hopefully converges to a (local) minimum point by means of the iteration

$$x_{i+1} = x_i - \alpha_i f'(x_i), \tag{60.2}$$

where α_i is a positive parameter. Since $\alpha_i > 0$, if $f'(x_i) > 0$ then $x_{i+1} < x_i$, and if $f'(x_i) < 0$ then $x_{i+1} > x_i$. This means that if $f'(x_i) > 0$, so that $f(x)$ is increasing at $x = x_i$, then taking $x_{i+1} < x_i$ should result in $f(x_{i+1}) < f(x_i)$, and thus x_{i+1} should be closer to a minimum point than x_i. A similar argument applies in the case $f'(x_i) < 0$.

It is clear that the choice of the parameter α_i is important. If α_i is too small, then the convergence will be slow, and if α_i is too large, the sequence x_i may start to oscillate.

Note that we may view the gradient method (60.2) for minimization of $f(x)$ as Fixed Point Iteration for computing a root of $f'(x) = 0$.

If steepest descent leads to the boundary Γ of Ω, then we may replace the steepest descent iteration by the *projected gradient method* defined

$$x_{i+1} = x_i - \alpha_i P f'(x_i),$$

where $Pf'(x_i)$ is the projection of $f'(x_i)$ onto the tangent plane to Γ at $x_i \in \Gamma$.

The general idea is thus to find roots of $f'(x) = 0$ using steepest descent for the minimization of $f(x)$ or equivalently fixed point iteration for $f'(x) = 0$. Once the roots of $f'(x) = 0$ have been determined, the minimization is reduced to a search on the boundary of Ω and the interior zeros of $f'(x)$.

60.8 Existence of a minimum value and a minimum point

We return to proof of the fundamental result which says that if $f : \Omega \to \mathbb{R}$ is Lipschitz continuous and Ω is a closed and bounded domain of $\mathbb{R}^d$, then there is a minimum point $\bar{x} \in \Omega$ with corresponding minimum value $f(\bar{x})$. We carry out the proof in for $d = 1$ so that $\Omega = [a, b]$ is a bounded closed interval. The proof in the case $d > 1$ is similar.

We shall prove that a Lipschitz continuous function $f : [a, b] \to \mathbb{R}$ on a closed and bounded interval $[a, b]$ has a minimum point by "constructing" a minimum point using the Bisection algorithm. We shall see that the "construction" is controversial at one step. Trying to resolve this issue yields added insight into the nature of minimization algorithms.

Normally, the proof we present here is considered so "difficult" that it is given only in "advanced" senior undergraduate or beginning graduate courses. With our good preparation on the Bisection algorithm and the nature of real numbers, we can plunge into the proof, and we will see that it is "easy" up to the non-constructive aspects.

We first recall that the Lipschitz continuity of $f(x)$ and the fact that $[a, b]$ is bounded implies that $f : [a, b] \to \mathbb{R}$ is bounded from above and below. In particular, there is some $m \in \mathbb{R}$ such that

$$f(x) \geq m \quad \text{for all } x \in [a, b]. \tag{60.3}$$

We say that m is a *lower bound* of $f:[a,b]\to\mathbb{R}$ if (60.3) holds. Clearly, there are many lower bounds since if m is a lower bound, any number $\underline{m}<m$ is also a lower bound.

In the proof, we shall use the concept of *greatest lower bound* defined as follows: we say that $\overline{m}$ is a greatest lower bound of $f:[a,b]\to\mathbb{R}$ if

$$f(x)\geq\overline{m}\quad\text{for all } x\in[a,b] \tag{60.4}$$

$$\text{for all } M>\overline{m}\quad\text{there is some } x\in[a,b]\quad\text{such that } f(x)<M. \tag{60.5}$$

In words, $\overline{m}$ is a greatest lower bound of $f:[a,b]\to\mathbb{R}$ if $\overline{m}$ is a lower bound of $f:[a,b]\to\mathbb{R}$ and any number bigger than $\overline{m}$ is not a lower bound for $f:[a,b]\to\mathbb{R}$. The concept of greatest lower bound has played an important role in the development of Calculus during the 20th century.

The proof now proceeds in two steps:

Step 1: Existence of a greatest lower bound $\overline{m}$ of $f:[a,b]\to\mathbb{R}$

We shall prove the existence of a greatest lower bound $\overline{m}$ by using the Bisection method. Let m be a lower bound of $f:[a,b]\to\mathbb{R}$, whose existence was established above. Set $y_0=m$ and $Y_0=f(b)$ and define $\hat{y}_1=\frac{1}{2}(y_0+Y_0)=\frac{1}{2}(m+f(b))$. Note that $y_0\leq\hat{y}_1\leq Y_0$. If $f(x)\geq\hat{y}_1$ for all $x\in[a,b]$, then set $y_1=\hat{y}_1$ and $Y_1=Y_0$. If not, then there is an $x\in[a,b]$ such that $f(x)<\hat{y}_1$, and we set $y_1=m$ and $Y_1=\hat{y}_1$. We have now passed from the pair (y_0,Y_0), or interval (y_0,Y_0), to the interval (y_1,Y_1). By construction, $f(x)\geq y_i$ for all $x\in[a,b]$ and $i=0,1$ and there is some $x\in[a,b]$ such that $f(x)<Y_i$ unless Y_0 or Y_1 is already a greatest lower bound.

Repeating this process, we get two sequences $\{y_i\}$ and $\{Y_i\}$ such that for $i=0,1,2,...,,$

$$\begin{gathered}
y_i<Y_i,\quad y_{i+1}\geq y_i\quad Y_{i+1}\leq Y_i,\\
0<Y_i-y_i=2^{-i}(Y_0-m),\\
f(x)\geq y_i\quad\text{for all } x\in[a,b],\\
\text{there is an } x\in[a,b]\quad\text{such that } f(x)<Y_i,
\end{gathered}$$

or some Y_i is a greatest lower bound. As in Chapter $\sqrt{2}$, we see that the sequences $\{y_i\}$ and $\{Y_i\}$ are Cauchy sequences and both converge to one real number, which we denote by $\overline{m}$. The number $\overline{m}$ is the greatest lower bound of $f:[a,b]\to\mathbb{R}$ since $\overline{m}$ satisfies the following two conditions:

$$\begin{gathered}
(f(x)\geq\overline{m}\quad\text{for all } x\in[a,b],\\
\text{for any } M>\overline{m} \text{ there is an } x\in[a,b]\quad\text{such that } f(x)<M.
\end{gathered}$$

We have now proved the existence of a greatest lower bound to the Lipschitz continuous function $f:[a,b]\to\mathbb{R}$ on the closed and bounded interval $[a,b]$. Note that this result also holds if (a,b) is a bounded open interval. We have thus not yet used the fact that $[a,b]$ is closed.

Step 1: Existence of a minimum point

We now construct a convergent sequence $\{x_i\}$ with $x_i \in [a, b]$ and

$$\lim_{i\to\infty} f(x_i) = \overline{m}.$$

Setting $\bar{x} = \lim_{i\to\infty} x_i$, we have $f(\bar{x}) = \overline{m}$ and thus $\bar{x}$ is a minimum point and we are done.

To construct $\{x_i\}$ we again use the Bisection algorithm as follows: set $x_0 = a$, and $X_0 = b$, and define $\hat{x}_1 = \frac{1}{2}(x_0 + X_0)$. If $f(x) > \overline{m}$ for all x such that $\hat{x}_1 < x \le X_1$, then we set $x_1 = x_0$ and $X_1 = \hat{x}_1$. If not, we set $x_1 = \hat{x}_1$ and $X_1 = X_0$. Repeating the process, we obtain a convergent sequence $\{x_i\}$ with limit $\bar{x}$ and by construction we have $f(\bar{x}) = \overline{m}$. Note that to guarantee that $\overline{x} \in [a, b]$, we need $[a, b]$ to be closed. We note that the minimum value (of course) is equal to the greatest lower bound.

We summarize in the following theorem:

Theorem 60.3 (Existence of minimum point) *Suppose $f : I \to \mathbb{R}$ is Lipschitz continuous and $I = [a, b]$ is a closed and bounded interval. Then there is a point $\bar{x} \in [a, b]$, where $f : I \to \mathbb{R}$ assumes a minimal value $\bar{m}$, that is, $f(x) \ge \bar{m}$ for all $x \in [a, b]$, and $f(\bar{x}) = \bar{m}$.*

In the proof of this theorem we used the Bisection algorithm twice. Setting $y = f(x)$, we may say that we first used the Bisection algorithm in the variable y to prove existence of a greatest lower bound $\overline{m}$ and then in the variable x to prove existence of a minimum point $\bar{x}$ satisfying $f(\bar{x}) = \overline{m}$.

60.9 Existence of greatest lower bound to a set bounded below

If we examine the proof of existence of a greatest lower bound to the Lipschitz continuous function $f : I \to \mathbb{R}$, we see that the crucial fact behind the proof is that $f : [a, b] \to \mathbb{R}$ is bounded below, that is there is a real number m such that $f(x) \ge m$ for all $x \in [a, b]$. We can interpret this in terms of a property of the range $R(f) = \{y : y = f(x) \text{ for some } x \in D(f) = [a, b]\}$, namely

$$y \ge m \quad \text{for all } y \in R(f).$$

This says that the set $R(f)$ is *bounded below.*

More generally, we say that a set A of real numbers is bounded from below if there is a real number m such that $y \ge m$ for all $y \in A$. Using the same argument as just used in the case $A = R(f)$, we obtain the following fundamental property of real numbers.

Theorem 60.4 (Existence of greatest lower bound) *Suppose A is a set of real numbers which is bounded from below, that is, there is a real* *number m such that $x \geq m$ for $x \in A$. Then the set A has a greatest lower bound $\overline{m} \in \mathbb{R}$ satisfying $x \geq \overline{m}$ for all $x \in A$ and for all $M > \overline{m}$ there is an $x \in A$ such that $x < M$.*

60.10 Analysis of the constructibility of minimum value and minimum point

We now discuss to what extent the above existence proof is constructive. There are two issues: (i) construction of the greatest lower bound, which is the same as the minimal value, and (ii) construction of a minimal point.

In the application of the Bisection algorithm in (i), we have to check if

$$f(x) \geq \hat{y}_i \quad \text{for all } x \in [a, b],$$

while in the application in (ii), we have to check if

$$f(x) > \overline{m} \text{ for all } x \text{ such that } \hat{x}_1 < x \leq X_1.$$

Both checks appear to involve *infinitely many* values of x. In the worst case this would require infinitely many comparisons. The number may be reduced if $f(x)$ is differentiable by using information concerning $f'(x)$. For example, the sign of $f'(x)$ indicates if $f(x)$ is increasing or decreasing which may be used to reduce the amount of comparison.

Thus, depending on the nature of the given function $f : I \to \mathbb{R}$, the given proof of existence of a minimum value and minimum point may be more or less constructive in nature.

Is it possible to make the proof fully constructive? We expect this to be possible if we accept to determine the minimum value up to a tolerance $TOL > 0$. Suppose then that the function $f(x)$ is Lipschitz continuous with Lipschitz constant L. We can then reduce all comparisons to a discrete grid of points of mesh size $\frac{1}{L}TOL$ between neighboring points.

To sum up, if $f : I \to \mathbb{R}$ is Lipschitz continuous and $[a, b]$ is bounded, then it is possible to determine the minimum value $f : I \to \mathbb{R}$ up to a given tolerance with a finite number of operations.

To determine an interior minimum point amounts to finding a root of $f'(x) = 0$ and thus the constructibility of a minimum point can be reduced to the constructibility of a root of $f'(x) = 0$. We discussed the cost of computing roots in Chapters *Fixed Point Iteration* and *Newton's method.*

60.11 A decreasing bounded sequence converges!

Suppose $\{x_i\}$ is a bounded decreasing sequence, that is $x_1 \geq x_2 \geq \cdots \geq x_n \geq x_{n+1} \geq \ldots$, and $x_n \geq m$ for all n for some number m. Then the set of all numbers x_n is bounded below, and thus has a greatest lower bound $\bar{m}$. We shall prove that $\lim_{n\to\infty} x_n = \bar{m}$. By the definition of greatest lower bound, for all $\epsilon > 0$ there is an x_N such that $\bar{m} \leq x_N \leq \bar{m} + \epsilon$. Since $x_n \leq x_N$ for $n \geq N$, and $x_n \geq \bar{x}$, it follows that $\bar{m} \leq x_n \leq \bar{m} + \epsilon$ for all $n \geq N$, which proves the desired result. We summarize in the following theorem which is a cornerstone of the analysis of functions of a real variables.

Theorem 60.5 *Suppose $\{x_i\}_{i=1}^{\infty}$ is a decreasing sequence that is bounded below or an increasing sequence that is bounded above. Then $\{x_i\}_{i=1}^{\infty}$ is convergent.*

Chapter 60 Problems

60.1. Find the maximal and minimal values of the function $f(x_1, x_2) = x_1^2 + 2x_2^2 - x_1$ on the unit disc $x_1^2 + x_2^2 \leq 1$.

60.2. Find the point of the plane $3x_1 + 4x_2 - x_3 = 26$ which is closest to the origin.

60.3. Find the shape of a box (with top included) which for given surface area has maximal volume.

60.4. Seek minimum and maximum values of the following functions: (a) $f(x_1, x_2) = (1 + x_1^2 + x_2^2)^{-1}$ for $(x_1, x_2) \in \mathbb{R}^2$, (b) $f(x_1, x_2) = x_1x_2$ for $x_1^2 + x_2^2 \leq 1$, (c) $f(x_1, x_2, x_3) = x_1 + x_2 + x_3$ for $x_1^2 + x_2^2 + x_3^2 \leq 1$.

60.5. Show that the function $x_1^4 + x_2^4 + x_3^4 - 4x_1x_2x_3$ has a minimum point at $(x_1, x_2.x_3) = (1, 1, 1)$.

60.6. Find the triangle of largest area that can be inscribed in a given circle.

60.7. Find the point on the curve $x_2 = x_1^2$ which is closest to the point $(0, 1)$.

60.8. Determine the constants a_0 and a_1 which minimize for a given function $f : [0, 1] \to \mathbb{R}$, the integral

$$\int_0^1 (f(x) - a_0 - a_1x)^2\, dx.$$

60.9. Find the maximum value of $x_1 + x_2 + ... + x_n$ subject to the condition $x_1^2 + x_2^2 + ... + x_n^2 \leq 1$.

60.10. A *stationary point* of a function $f : \mathbb{R}^n \to \mathbb{R}$ is a point $x \in \mathbb{R}^n$ such that $f'(x) = 0$. Determine if any of the stationary points of the following functions is a maximum or minimum point: (a) $f(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2 - x_1 - x_2 + x_3 + 1$, (b) $f(x_1, x_2, x_3) = x_1^2 + x_2^2 + 2x_3^2 + 4x_1 - x_2 + x_3 + 5$, (c) $f(x_1, x_2, x_3) = \cos(x_1) + \cos(x_2) + \cos(x_3)$.

61
The divergence, rotation and Laplacian

.. Stokes was a very important formative influence on subsequent generations of Cambridge men, including Maxwell. With Green, who in turn had influenced him, Stokes followed the work of the French, especially Lagrange, Laplace, Fourier, Poisson and Cauchy. This is seen most clearly in his theoretical studies in optics and hydrodynamics; but it should also be noted that Stokes, even as an undergraduate, experimented incessantly. Yet his interests and investigations extended beyond physics, for his knowledge of chemistry and botany was extensive, and often his work in optics drew him into those fields. (Parkinson)

Appointed professor of mathematics at the Ecole Polytechnique in 1809 Ampére held posts there until 1828. Ampére and Cauchy shared the teaching of analysis and mechanics and there was a great contrast between the two with Cauchy's rigorous analysis teaching leading to great mathematical progress but found extremely difficult by students who greatly preferred Ampére's more conventional approach to analysis and mechanics. (O'Connor and Robertson)

61.1 Introduction

We saw previously that the gradient of a function of several variables is a practically useful *differential operator*. In this chapter, we introduce some other useful operators, including the *divergence*, *rotation* and the *Laplacian*, together with the gradient play a fundamental role in mathematical modeling in science and engineering. We first define the operators in $\mathbb{R}^2$ and then in $\mathbb{R}^3$, noting that the rotation takes somewhat different forms in $\mathbb{R}^2$ and $\mathbb{R}^3$.

FIGURE 61.1. Napoleon to Laplace (1749-1827): "You have written this huge book on the system of the world without once mentioning the Author of the Universe". Laplace to Napoleon: "Sire, I had no need of this hypothesis"

61.2 The gradient, divergence, rotation and Laplacian in $\mathbb{R}^2$

We recall that the *gradient* of a function $u : \mathbb{R}^2 \to \mathbb{R}$, denoted grad u or ∇u, is the vector-valued function formed by the first order partial derivatives of u, i.e.

$$\text{grad } u = \nabla u = \left(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}\right).$$

The *divergence* of a vector function $u = (u_1, u_2) : \mathbb{R}^2 \to \mathbb{R}^2$, denoted $\operatorname{div} u$ or $\nabla \cdot u$, is the scalar function defined by

$$\operatorname{div} u = \nabla \cdot u = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2}.$$

Formally, we have

$$\nabla \cdot u = (\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}) \cdot (u_1, u_2)$$

where we may think of $(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2})$ "as a vector" and let the dot indicate a "scalar product". This idea applies to all the formulas below involving ∇ combined with the operators $\cdot$ and $\times$.

The *rotation* of a vector function $u : \mathbb{R}^2 \to \mathbb{R}^2$, denoted by rot u or $\nabla \times u$, is the scalar function

$$\operatorname{rot} u = \nabla \times u = \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2} = (\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}) \times (u_1, u_2).$$

If $u : \mathbb{R}^2 \to \mathbb{R}$ is a scalar function, then $\operatorname{rot} u = \nabla \times u$ is defined as the vector function

$$\operatorname{rot} u = \nabla \times u = \left(\frac{\partial u}{\partial x_2}, -\frac{\partial u}{\partial x_1}\right).$$

The different appearances of $\operatorname{rot} u = \nabla \times u$, with u a scalar or $u = (u_1, u_2)$ a vector function will be explained when we pass to $\mathbb{R}^3$ below. For now, it may be helpful to recall the different appearances of $a \times b$ with $a, b \in \mathbb{R}^2$ or $a, b \in \mathbb{R}^3$.

The following identities follow directly from the definitions for any function u:

$$\begin{aligned} \nabla \cdot (\nabla \times u) &= \text{div (rot } u) = 0, \qquad (u : \mathbb{R}^2 \to \mathbb{R}^2) \\ \nabla \times (\nabla u) &= \text{rot (grad } u) = 0, \qquad (u : \mathbb{R}^2 \to \mathbb{R}). \end{aligned} \tag{61.1}$$

Finally, the *Laplacian* Δu of a function $u : \mathbb{R}^2 \to \mathbb{R}$ is defined by

$$\Delta u = \nabla \cdot (\nabla u) = \text{div (grad } u) = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2},$$

where $\frac{\partial^2 u}{\partial x_i^2} = \frac{\partial}{\partial x_i}(\frac{\partial u}{\partial x_i})$.

61.3 The Laplacian in $\mathbb{R}^2$ in polar coordinates

In polar coordinates $x = (x_1, x_2) = (r\cos(\theta), r\sin(\theta))$ with $r \geq 0$ and $0 \leq \theta < 2\pi$, the Laplacian takes the form

$$\Delta u = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}. \tag{61.2}$$

This follows by a routine computation using that the Jacobian of the mapping $x = (r\cos(\theta), r\sin(\theta))$, in the notation (54.9) is given by

$$\frac{d(x_1, x_2)}{d(r, \theta)} = \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix},$$

so that

$$\frac{d(r, \theta)}{d(x_1, x_2)} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta)/r & \cos(\theta)/r \end{pmatrix},$$

and thus by the Chain rule

$$\frac{\partial}{\partial x_1} = \cos(\theta)\frac{\partial}{\partial r} - \frac{\sin(\theta)}{r}\frac{\partial}{\partial \theta} \quad \text{and} \quad \frac{\partial}{\partial x_2} = \sin(\theta)\frac{\partial}{\partial r} + \frac{\cos(\theta)}{r}\frac{\partial}{\partial \theta}.$$

61.4 Basic examples with $\nabla \cdot u = 1$, $\nabla \times v = 1$ and $\Delta w = 1$

The function $u : \mathbb{R}^2 \to \mathbb{R}^2$ given by $u(x) = \frac{1}{2}(x_1, x_2)$, satisfies

$$\nabla \cdot u(x) = 1.$$

The function $v : \mathbb{R}^2 \to \mathbb{R}^2$ given by $v(x) = \frac{1}{2}(-x_2, x_1)$, satisfies

$$\nabla \times v(x) = 1.$$

The function $w : \mathbb{R}^2 \to \mathbb{R}$ given by $w(x) = \frac{1}{4}(x_1^2 + x_2^2)$, satisfies

$$\Delta w = 1.$$

We plot these basic examples in Fig. 61.2

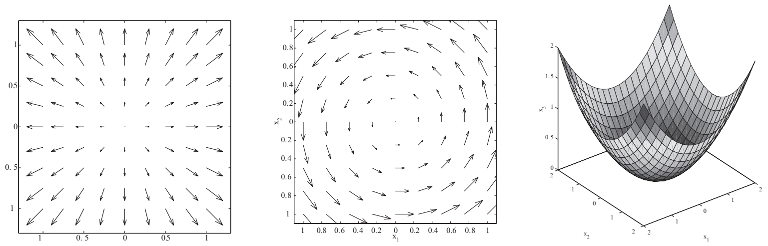


FIGURE 61.2. Basic examples satisfying $\nabla \cdot u = 1$, $\nabla \times v = 1$ and $\Delta w = 1$, respectively.

We see that $u(x)$ "explodes", $v(x)$ "rotates" and $w(x)$ is a "hump".

61.5 Invariance of the Laplacian under rigid coordinate transformations

It follows from the form of the Laplacian in polar coordinates, that the Laplacian is invariant under rotations and translations in $\mathbb{R}^2$, i.e. so-called *rigid transformations* of the form

$$\begin{aligned}\tilde{x}_1 &= \cos(\alpha)x_1 + \sin(\alpha)x_2 + a_1\\ \tilde{x}_2 &= -\sin(\alpha)x_1 + \cos(\alpha)x_2 + a_2,\end{aligned}$$

where (x_1, x_2) are the old coordinates and $(\tilde{x}_1, \tilde{x}_2)$ the new ones. In other words, the Laplacian takes exactly the same form in the two coordinate systems:

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = \frac{\partial^2 u}{\partial \tilde{x}_1^2} + \frac{\partial^2 u}{\partial \tilde{x}_2^2}.$$

This fact is reflected in the observation that the Laplace operator typically occurs in *isotropic* models that have the same properties in all directions.

61.6 The gradient, divergence, rotation and Laplacian in $\mathbb{R}^3$

The *gradient* of a function $u : \mathbb{R}^3 \to \mathbb{R}$, denoted grad u or ∇u, is the vector-valued function formed by the set of first order partial derivatives of u, i.e.

$$\text{grad } u = \nabla u = \left(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, \frac{\partial u}{\partial x_3}\right).$$

For a vector function $u : \mathbb{R}^3 \to \mathbb{R}^3$, the divergence div u is a scalar function defined by

$$\text{div } u = \sum_{i=1}^{3} \frac{\partial u_i}{\partial x_i},$$

and $\text{rot}\, u$ is the vector function

$$\text{rot } u = \nabla \times u = \left(\frac{\partial u_3}{\partial x_2} - \frac{\partial u_2}{\partial x_3}, \frac{\partial u_1}{\partial x_3} - \frac{\partial u_3}{\partial x_1}, \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}\right).$$

We now explain the relation of the operator of rotation $\nabla\times$ in $\mathbb{R}^3$ to the operator of rotation $\nabla\times$ in $\mathbb{R}^2$ introduced above. Consider first a function $u : \mathbb{R}^3 \to \mathbb{R}^3$ of the form $u = (u_1, u_2, 0)$ with u_1 and u_2 being independent of x_3 so that effectively $u_i : \mathbb{R}^2 \to \mathbb{R}$ with $u_i = u_i(x_1, x_2)$ for $i = 1, 2$.. We have

$$\nabla \times u = (0, 0, \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}) = (0, 0, \nabla \times (u_1, u_2)).$$

Secondly, if $u : \mathbb{R}^3 \to \mathbb{R}^3$ has the form $u = (0, 0, u_3)$ with u_3 independent of x_3, so that effectively $u_3 : \mathbb{R}^2 \to \mathbb{R}$, then

$$\nabla \times u = (\frac{\partial u_3}{\partial x_2}, -\frac{\partial u_3}{\partial x_1}, 0) = (\nabla \times u_3, 0).$$

We conclude that $\nabla \times u$ for $u : \mathbb{R}^2 \to \mathbb{R}$ and $\nabla \times u$ for $u : \mathbb{R}^2 \to \mathbb{R}^2$, may be viewed as special cases of $\nabla \times u$ for $u : \mathbb{R}^3 \to \mathbb{R}^3$.

The *Laplacian* Δu of a function $u : \mathbb{R}^2 \to \mathbb{R}$ is defined by

$$\Delta u = \nabla \cdot (\nabla u) = \text{div (grad } u) = \sum_{i=1}^{3} \frac{\partial^2 u}{\partial x_i^2}.$$

By direct computation we verify the following identities:

$$\begin{aligned} \nabla \cdot (\nabla \times u) &= 0, \\ \nabla \times (\nabla u) &= 0, \\ \nabla \times (\nabla \times u) &= -\Delta u + \nabla(\nabla \cdot u). \end{aligned} \tag{61.3}$$

61.7 Basic examples with $\nabla \cdot u = 1$, $\nabla \times v = (0, 0, 1)$ and $\Delta w = 1$

The function $u : \mathbb{R}^3 \to \mathbb{R}^3$ given by $u(x) = \frac{1}{3}x$, satisfies

$$\nabla \cdot u(x) = 1.$$

The function $v : \mathbb{R}^3 \to \mathbb{R}^3$ given by $v(x) = \frac{1}{2}(-x_2, x_1, 0)$, satisfies

$$\nabla \times v(x) = (0, 0, 1).$$

The function $w : \mathbb{R}^3 \to \mathbb{R}$ given by $w(x) = \frac{1}{6}\|x\|^2$, satisfies

$$\Delta w = 1.$$

We plot these basic examples in Fig. 61.3

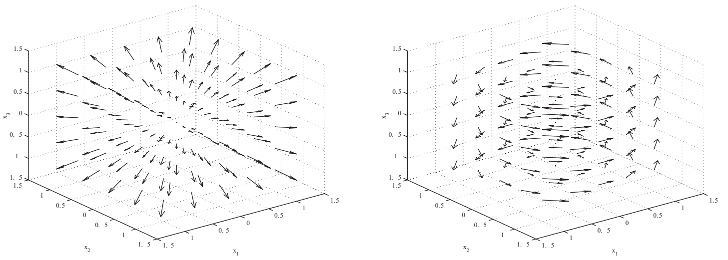


FIGURE 61.3. Basic examples in $\mathbb{R}^3$ satisfying $\nabla \cdot u = 1$, $\nabla \times v = 1$.

We see again that $u(x)$ "explodes", $v(x)$ "rotates" along the x_3 axis while the "hump" $w(x)$ is difficult to visualize.

61.8 The Laplacian in spherical coordinates

In *spherical coordinates*.

$$x = (x_1, x_2, x_3) = (r\sin(\varphi)\cos(\theta), r\sin(\varphi)\sin(\theta), r\cos(\varphi)),$$

where $r \geq 0$, $0 \leq \theta < 2\pi$ and $0 \leq \varphi < \pi$, the Laplacian is given by

$$\Delta u = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right) + \frac{1}{r^2\sin(\theta)}\frac{\partial}{\partial \theta}\left(\sin(\theta)\frac{\partial u}{\partial \theta}\right) + \frac{1}{r^2\sin^2(\theta)}\frac{\partial^2 u}{\partial \varphi^2}. \quad (61.4)$$

The Laplacian is invariant under orthogonal coordinate transformations in $\mathbb{R}^3$.

EXAMPLE 61.1. Consider the velocity field generated by rotation around a vector $\omega \in \mathbb{R}^3$ with angular speed $\|\omega\|$, that is the vector field

$$v(x) = \omega \times x.$$

We compute

$$\begin{aligned}\nabla \times v(x) &= \nabla \times (\omega_2 x_3 - \omega_3 x_2, \omega_3 x_1 - \omega_1 x_3, \omega_1 x_2 - \omega_2 x_1)\\ &= (2\omega_1, 2\omega_2, 2\omega_3) = 2\omega.\end{aligned}$$

We conclude that the rotation $\nabla \times v(x)$ of a velocity field $v(x)$ generated by a rotation according to a given vector ω is equal to 2ω. This motivates the name of the differential operator $\nabla\times$ as the "rotation".

EXAMPLE 61.2. A basic formula of electromagnetics expressing Ampère's law states that the *magnetic field* H generated by a unit electrical current flowing through the x_3-axis in the positive direction, is given by

$$H(x) = H(x_1, x_2, x_3) = \frac{1}{2\pi}\frac{(-x_2, x_1, 0)}{x_1^2 + x_2^2} \quad \text{for } x_1^2 + x_2^2 > 0. \qquad (61.5)$$

We compute

$$\nabla \times H(x) = \frac{1}{2\pi}(0, 0, \frac{\partial}{\partial x_1}\frac{x_1}{x_1^2 + x_2^2} - \frac{\partial}{\partial x_2}\frac{-x_2}{x_1^2 + x_2^2}) = 0 \quad \text{for } x_1^2 + x_2^2 > 0.$$

Thus $\nabla \times H(x) = 0$ for $x_1^2 + x_2^2 > 0$, which is just *Amperes's Law* $\nabla \times H = J$, where J is the current density, noting that $J(x) =$ for $x_1^2 + x_2^2 > 0$, i.e. outside the x_3-axis. Amperes's Law is one of *Maxwell's equations.* Below we shall show how to interpret the equation $\nabla \times H(x) = J(x)$ for $x_1^2 + x_2^2 = 0$ and motivate the factor $\frac{1}{2\pi}$ in (61.5).

Chapter 61 Problems

61.1. Let $F = (5x_1 - 3x_1x_2 + x_3^2, \sin(x_1)\cos(x_1) + x_1, \sin(x_1)\exp(x_1x_2))$. With $x = (1, 2, 3)$, compute (a) $\nabla \cdot F$, (b) $\nabla \times F$, (c) $\nabla(\nabla \cdot F)$, (d) $\nabla \times (\nabla \times F)$.

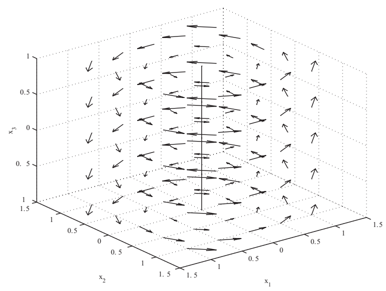


FIGURE 61.4. The magnetic field around a current through the x_3-axis

61.2. Interpret the expression $(\nabla \times \nabla)u$ in a reasonable way and show that $(\nabla \times \nabla)u = 0$ for any u. Compare with $\nabla \times (\nabla \times u)$.

61.3. Show that for appropriate function u and v

1. $\nabla(uv) = (\nabla u)v + u(\nabla v)$,
2. $\nabla \cdot (uv) = (\nabla u) \cdot v + u(\nabla \cdot v)$,
3. $\nabla \times (uv) = (\nabla u) \times v + u(\nabla \times v)$,
4. $\nabla \cdot (u \times v) = v \cdot (\nabla \times u) - u \cdot (\nabla \times v)$,
5. $\nabla \times (u \times v) = (v \cdot \nabla)u - (\nabla \cdot u)v - (u \cdot \nabla)v + (\nabla \cdot v)u$,
6. $\nabla(u \cdot v) = (u \cdot \nabla)v + (v \cdot \nabla)u + u \times (\nabla \times v) + v \times (\nabla \times u)$.

61.4. Compute $\nabla(r \cdot F(r))$ where $r = \|x\|$.

61.5. Prove that the velocity field $v(x) = \omega \times x$, where $\omega \in \mathbb{R}^3$ is a given vector, satisfies $\nabla \cdot v(x) = 0$. Interpret the result in fluid mechanical terms.

61.6. Prove directly using the Chain rule that the Laplacian in $\mathbb{R}^2$ and $\mathbb{R}^3$ is invariant under rigid coordinate transformations.

61.7. Prove (61.3), (61.2) and (61.4).

61.8. Show that if $u : \mathbb{R}^2 \to \mathbb{R}$, then $\nabla \times (\nabla \times u) = \text{rot (rot u)} = -\Delta u$.

61.9. Show that the function $u : \mathbb{R}^2 \to \mathbb{R}$ given by $u(x) = c_1 \log(\|x\|) + c_2$ with c_1 and c_2 constants, is a solution of the Laplace equation $\Delta u(x) = 0$ in $\mathbb{R}^2$ for $x \neq 0$.

61.10. Prove that the function $u : \mathbb{R}^3 \to \mathbb{R}$ given by $u(x) = c_1\|x\|^{-1} + c_2$, with c_1 and c_2 constants, is a solution of Laplace's equation $\Delta u(x) = 0$ in $\mathbb{R}^3$ for $x \neq 0$.

61.11. Show that the divergence is invariant under rigid coordinate transformations. Does the rotation have the same property?

> All the effects of Nature are only the mathematical consequences of a small number of immutable laws. (Laplace)

62
Meteorology and Coriolis forces*

> Any teacher who stands up in front of a class and says that Coriolis force determines which way the water flows from a sink or bathtub, should not only read Fraser's Bad Coriolis Web page (www.ems.psu.edu/ fraser/Bad/BadCoriolis.html), but be required to copy it on the blackboard 100 times. (Jack Williams, USA TODAY)

62.1 Introduction

A common weather map shows the level curves of the air pressure p, the so-called *isobars*. Intuition might suggest that the wind will blow from high pressure to low pressure, i.e. in the opposite direction to the pressure gradient ∇p and orthogonal to the isobars. However, this turns out to be completely false. In fact, the wind circles around a center of low pressure in a counter-clockwise direction on the North hemisphere and in a clockwise direction on the Southern hemisphere, and in the opposite directions around centers of high pressure. Thus the wind blows along the isobars, instead of orthogonal to the isobars. This fact is well-known to sailors, making it possible to easily and accurately predict the wind direction if the centers of the low and high pressures are known. The reason is that the Earth is rotating, which creates a force of acceleration called the *Coriolis force*. This causes the wind to deviate to the right on the Northern hemisphere and to the left on the Southern hemisphere (away from the equator). The effect is that the wind circles around a center of low pressure in a counter-clockwise direction on the Northern hemisphere, as any weather map in a newspaper indicates. The Coriolis force is felt on a turn-around when seeking to change position in the radial direction, which causes an (unexpected) force in the tangential direction.

62.2 A basic meteorological model

We shall now derive a simple model for the motion of the atmosphere, which predicts that the wind should revolve around centers of low and high pressure. The model takes the form

$$\nabla p = \rho 2\omega \times v, \tag{62.1}$$

where p is the pressure, v is the wind velocity, $\omega \in \mathbb{R}^3$ is the angular velocity of the Earth, and ρ is the density of the atmosphere. The quantity $2\omega \times v$ is an approximation of the Coriolis acceleration and the equation $\nabla p = \rho 2\omega \times v$ gives a balance of the pressure force ∇p and the Coriolis force $\rho 2\omega \times v$. Here ∇p represents the gradient in the plane of the surface of the Earth and the model applies to "caps" on the Northern or Southern away from the Equator, say above or below the 60 degree latitude, where we can approximate the surface of the Earth by a flat disc, see Fig. 62.1, that is, the "world" of the sailor and the wind is a big flat turn-around.

We see that (62.1) states that ∇p is orthogonal to the direction of the wind. If we know p, we can determine the wind direction and speed from (62.1).

62.3 Rotating coordinate systems and Coriolis acceleration

To derive the expression $2\omega \times v$ of the Coriolis force, we need to study coordinate transformations from one fixed coordinate system to a rotating coordinate system. We thus let $\{e_1, e_2, e_3\}$ be a fixed orthonormal reference coordinate system for $\mathbb{R}^3$, and we let $\{\bar{e}_1, \bar{e}_2, \bar{e}_3\}$ be another orthonormal coordinate system with the same origin, which rotates around the fixed vector $\omega \in \mathbb{R}^3$ with the angular speed $\|\omega\|$. More precisely, if $x(t) = x_1(t)e_1 + x_2(t)e_2 + x_3(t)e_3$ are the reference coordinates of a fixed point in the rotating coordinate system, then according to Fig. 62.2 we have

$$\frac{dx}{dt} = \omega \times x, \tag{62.2}$$

since $\frac{dx}{dt}$ is perpendicular to both ω and x, and $\|\frac{dx}{dt}\| = \|\omega\|\|x\|\sin(\theta)$, where $\theta \in [0, \pi]$ is the angle between ω and x. In particular we have for the basis

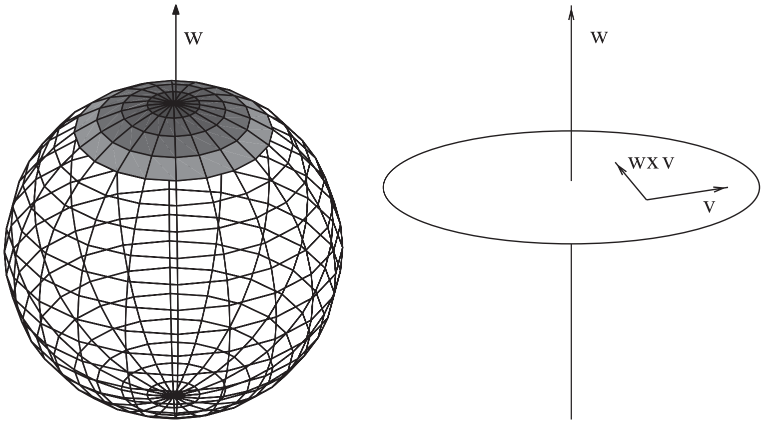


FIGURE 62.1. Northern Hemisphere **Change w to** ω

vectors of the moving coordinate system

$$\frac{d\bar{e}_i}{dt} = \omega \times \bar{e}_i, \quad i = 1, 2, 3. \tag{62.3}$$

Consider now a moving point with coordinates $x(t)$ in the fixed reference system and coordinates $\bar{x}(t)$ in the rotating system, so that

$$\begin{aligned} x(t) &= x_1(t)e_1 + x_2(t)e_2 + x_3(t)e_3, \\ \bar{x}(t) &= \bar{x}_1(t)\bar{e}_1(t) + \bar{x}_2(t)\bar{e}_2(t) + \bar{x}_3(t)\bar{e}_3(t), \end{aligned}$$

and of course $x(t) = \bar{x}(t)$. In particular, we may this way seek the coordinates of the basis vectors $\bar{e}_i(t)$ in the fixed system $\{e_1, e_2, e_3\}$. We now compute the velocity $\frac{dx}{dt}$ by differentiating $x(t) = \bar{x}(t)$ with respect to t to get

$$\begin{aligned} \frac{dx}{dt} &= \frac{d}{dt}\bar{x}(t) = \frac{d\bar{x}_1}{dt}\bar{e}_1 + \frac{d\bar{x}_2}{dt}\bar{e}_2 + \frac{d\bar{x}_3}{dt}\bar{e}_3 + \bar{x}_1\frac{d\bar{e}_1}{dt} + \bar{x}_2\frac{d\bar{e}_2}{dt} + \bar{x}_3\frac{d\bar{e}_3}{dt} \\ &= \frac{d\bar{x}_1}{dt}\bar{e}_1 + \frac{d\bar{x}_2}{dt}\bar{e}_2 + \frac{d\bar{x}_3}{dt}\bar{e}_3 + \bar{x}_1(\omega \times \bar{e}_1) + \bar{x}_2(\omega \times \bar{e}_2) + \bar{x}_3(\omega \times \bar{e}_3), \end{aligned}$$

where we used (62.3). We can write this expression as

$$\frac{dx}{dt} = \frac{\bar{d}\bar{x}}{dt} + \omega \times x, \tag{62.4}$$

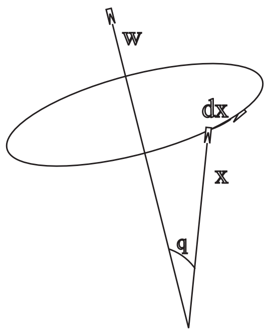


FIGURE 62.2. A vector x rotating with angular velocity ω.**Change w to ω and q to θ**

if we agree to write

$$\frac{\bar{d}\bar{x}}{dt} = \frac{d\bar{x}_1}{dt}\bar{e}_1 + \frac{d\bar{x}_2}{dt}\bar{e}_2 + \frac{d\bar{x}_3}{dt}\bar{e}_3.$$

The velocity of $x(t) = \bar{x}(t)$ in the fixed reference system is $\frac{dx}{dt}$, while $\frac{\bar{d}\bar{x}}{dt}$ is the velocity vs the rotating system involving the derivatives $\frac{\bar{d}}{dt}\bar{x}_i(t)$. In particular, if the point is fixed in the rotating system so that $\frac{\bar{d}\bar{x}}{dt} = 0$, then we retrieve (62.2) and (62.3).

We now seek a corresponding formula for the accelerations. We differentiate with respect to t once more, and using (62.4) with x replaced by $\frac{\bar{d}\bar{x}}{dt}$, we get

$$\begin{aligned}\frac{d^2x}{dt^2} &= \frac{d}{dt}\big(\frac{\bar{d}\bar{x}}{dt} + \omega \times x\big) = \frac{d}{dt}\big(\frac{\bar{d}\bar{x}}{dt}\big) + \omega \times \frac{dx}{dt}\\ &= \frac{\bar{d}}{dt}\big(\frac{\bar{d}\bar{x}}{dt}\big) + \omega \times \frac{\bar{d}\bar{x}}{dt} + \omega \times \big(\frac{\bar{d}\bar{x}}{dt} + \omega \times x\big).\end{aligned}$$

We can write this as

$$\frac{d^2x}{dt^2} = \frac{\bar{d}^2\bar{x}}{dt^2} + 2\omega \times \frac{\bar{d}\bar{x}}{dt} + \omega \times \big(\omega \times x\big). \tag{62.5}$$

Here, $\omega \times \big(\omega \times x\big)$ represents the *centripetal acceleration* and $2\omega \times \frac{\bar{d}\bar{x}}{dt}$ the *Coriolis acceleration*, and $\frac{d^2x}{dt^2}$ is the acceleration vs the reference system and $\frac{\bar{d}^2\bar{x}}{dt^2}$ the acceleration vs the rotating system.

By Newton's Law $F = ma$, *acceleration* is directly coupled to *force*, and thus both the centripetal and the Coriolis acceleration show up as forces in the fixed reference system. Both these forces in fact have a somewhat mysterious character; we have through massive daily experience become quite familiar with the centripetal acceleration, while the Coriolis force still presents surprises to most of us.

If the rotation speed $\|\omega\|$ is relatively small, then we can neglect the centripetal acceleration and we get

$$\frac{d^2x}{dt^2} \approx \frac{\bar{d}^2\bar{x}}{dt^2} + 2\omega \times \frac{\bar{d}\bar{x}}{dt}, \tag{62.6}$$

which leads to the model (62.1). Note that we use the rotating coordinate system in our "world", and thus $\frac{\bar{d}\bar{x}}{dt}$ is the relevant velocity.

Chapter 62 Problems

62.1. Motivate (62.1) using (62.6).

62.2. Inspect the isobars of a weather map and compute wind direction from (62.1) and compare with the wind direction of the map.

62.3. Study the effect of the Coriolis acceleration at the Equator.

62.4. Show that the centripetal acceleration of a body moving in a circle with radius r with speed v is equal to $\frac{v^2}{r}$.

62.5. The Gulf Stream is the reason Scandinavia is not deep frozen like Alaska. Explain why the Gulf Stream bends over from North America to North Europe.

62.6. Consider a car driving East-West along a certain latitude. At what speed is the Coriolis force on the car of the same size as the centripetal force? Determine this speed as a function of the latitude and find out at which latitudes the minimum and maximum is attained.

62.7. A bucket of water is spinning around its center with angular velocity ω. What is the shape of the water surface?

62.8. A pendulum of length l swings back and forth once every period of length $t = \sqrt{l/g}$, where g is the acceleration of gravity. Compute the Coriolis force on the pendulum at latitude θ (i.e. at an angle θ from the equator). This Coriolis force makes the plane in which the pendulum swings rotate, i.e. if the pendulum swings north–south at one instant, it will later swing west–east. Find the time T after which the pendulum swings in the initial direction once again as function of the latitude. What is the period on you latitude?

63
Curve integrals

> We can scarcely believe that Ampère really discovered the law of action by means of the experiments which he describes. We are led to suspect, what, indeed, he tells us himself, that he discovered the law by some process which he has not shown us, and that when he had afterwards built up a perfect demonstration he removed all traces of the scaffolding by which he had raised it. (Maxwell about Ampères *Memoir on the Mathematical Theory of Electrodynamic Phenomena, Uniquely Deduced from Experience*)

63.1 Introduction

In this chapter we introduce the concept of an *integral over a curve* or *curve integral*, and develop some applications including *arc length*, *work* and *line integrals*. We start with plane curves parameterized by functions $s : I \to \mathbb{R}^2$, where $I = [a, b]$ is an interval of the real line $\mathbb{R}$. We then generalize to curves in $\mathbb{R}^n$ parameterized by functions $s : I \to \mathbb{R}^n$ with $n \geq 2$.

63.2 The length of a curve in $\mathbb{R}^2$

Let Γ be a curve in $\mathbb{R}^2$ given by the function $s : I \to \mathbb{R}^2$, where $I = [a, b]$ is an interval of $\mathbb{R}$, that is, $\Gamma = \{s(t) \in \mathbb{R}^2 : t \in I\}$, or $\Gamma = s(I)$, see Fig. 63.1.

We now try to determine the *length* of Γ. We shall see that this leads to the introduction of the notion of an integral over a curve or a curve integral.

To define the length of a curve, we view the curve Γ as the being made up of little pieces of Γ. If the little pieces are sufficiently small, we can get away with approximating them by straight segments, and the length of a straight piece of curve is easy to compute. To find the total length of Γ, we will sum the lengths of all the little pieces forming Γ. We will find the integral is useful for this purpose.

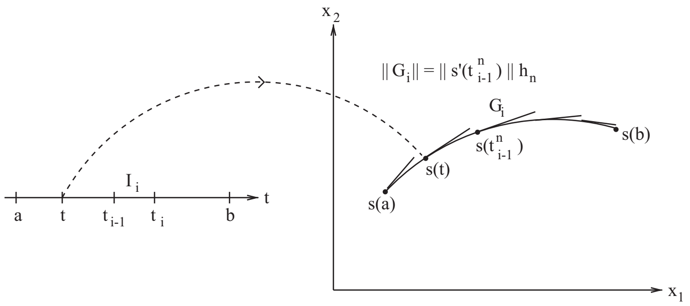


FIGURE 63.1. The total length of curve is the sum of the lengths of little pieces of the curve.

Let $a = t_0 < t_1 < \dots < t_n = b$ be a subdivison of I into intervals $I_i = (t_{i-1}, t_i]$. Consider the following linear approximation of the mapping $s(t)$ restricted to the subinterval I_i, see Fig. 63.1,

$$\bar{s}(t) = s(t_{i-1}) + (t - t_{i-1})s'(t_{i-1}).$$

The mapping $\bar{s}$ maps I_i onto the line segment Γ_i of length

$$\|s'(t_{i-1})\|(t_i - t_{i-1}),$$

and it is thus natural to use

$$L_n(\Gamma) = \sum_{i=1}^{n} \|s'(t_{i-1})\|(t_i - t_{i-1})$$

as an approximation of the length of Γ. Assuming that $\|s'(t)\|$ is Lipschitz continuous on I and assuming that $\max_i(t_i - t_{i-1})$ tends to zero as n tends to infinity, we can use the usual arguments to show that $\{L_n(\Gamma)\}_{n=1}^{\infty}$ is a Cauchy sequence and thus converges to a limit, which we denote by $L(\Gamma)$. We define this limit to be the *length* of Γ:

$$L(\Gamma) = \int_I \|s'(t)\|\, dt. \tag{63.1}$$

This formula expresses the length of a curve $\Gamma = s(I)$ as an integral over the parameter domain I of Γ with the modulus $\|s'(t)\|$ of the derivative of the representing function $s : I \to \mathbb{R}^2$ as a weight. Formally, we have $ds = \|s'(t)\|dt$, where ds represents the increase of the length of the curve corresponding to an increase dt of the parameter t; the function $\|s'(t)\|$ gives the local "change of scale" between the "element of curve length" ds; and the "parameter element" dt, see Fig. 63.1. We are thus led to write

$$L(\Gamma) = \int_\Gamma ds = \int_I \|s'(t)\|\, dt.$$

We will return to this notation in the next section.

EXAMPLE 63.1. We compute the length of the circumference Γ of a circle of radius 1 centered at the origin. The curve Γ is given by the function $s : [0, 2\pi) \to \mathbb{R}^2$ with $s(t) = (\cos(t), \sin(t))$ and $0 \le t < 2\pi$. We have $s'(t) = (-\sin(t), \cos(t))$ and $\|s'(t)\| = 1$, and thus

$$L(\Gamma) = \int_0^{2\pi} \|s'(t)\|\, dt = \int_0^{2\pi} dt = 2\pi.$$

We conclude that the the length of the circumference of a circle of radius 1 is equal to 2π (no big surprise). We check the result using a different parametrization. The upper semi-circle Γ_+ of Γ can be parameterized by $s : [-1, 1] \to \mathbb{R}^2$ given by $s(t) = (t, \sqrt{1-t^2})$ with $-1 \le t \le 1$. We have

$$s'(t) = (1, -\frac{t}{\sqrt{1-t^2}}), \quad \|s'(t)\| = \frac{1}{\sqrt{1-t^2}},$$

and thus

$$L(\Gamma) = 2L(\Gamma_+) = \int_{-1}^{1} \frac{1}{\sqrt{1-t^2}}\, dt = 2\big[\arcsin(t)\big]_{-1}^{1} = 2(\frac{\pi}{2} - (-\frac{\pi}{2})) = 2\pi.$$

63.3 Curve integral

Let $\Gamma = s(I)$ be a curve in $\mathbb{R}^2$ given by the function $s : I \to \mathbb{R}^2$, where $I = [a, b]$ is an interval of $\mathbb{R}$, and let $u : \Gamma \to \mathbb{R}$ be a function defined on Γ. We assume that the tangent $s' : I \to \mathbb{R}^2$ and the function $u : \Gamma \to \mathbb{R}$ are both Lipschitz continuous, which guarantees that $\|s'(t)\|$ and $u(s(t))$ are both Lipschitz continuous on I. We define the *integral of* u *over* Γ by

$$\int_\Gamma u\, ds \equiv \int_\Gamma u(x)\, ds(x) \equiv \int_a^b u(s(t))\|s'(t)\|\, dt. \tag{63.2}$$

Formally, we have $ds = ds(x) = \|s'(t)\|\, dt$, where $x = s(t)$.

EXAMPLE 63.2. If Γ is an interval $[a,b]$ on the x_1-axis given by $s(t) = (t,0)$, $a \le t \le b$, then $s'(t) = (1,0)$, $\|s'(t)\| = 1$, and

$$\int_\Gamma u\,ds = \int_a^b u(x_1,0)\,dx_1.$$

EXAMPLE 63.3. Let $\Gamma = s(I)$ be the semicircle given by $s(t) = (\cos(t),\sin(t))$, $0 \le t \le \pi$, and $u(x) = u(x_1,x_2) = x_1^2$. Using $\|s'(t)\| = 1$, we get

$$\int_\Gamma u\,ds = \int_0^\pi \cos^2(t)\,dt = \frac{1}{2}\int_0^\pi (1-\cos(2t))\,dt = \frac{\pi}{2}.$$

63.4 Reparametrization

An important observation is that the value of a curve integral is independent of the parameterization of the curve. To see this, consider two different parameterizations $s : [a,b] \to \Gamma$ and $\sigma : [c,d] \to \Gamma$ of a curve Γ in $\mathbb{R}^2$. Associate to each $\tau \in [c,d]$ the unique value $t \in [a,b]$ such that $s(t) = \sigma(\tau)$, which defines $t = t(\tau))$ as a function of τ (assuming that the curve does not cross itself), so that $\sigma(\tau) = s(t(\tau))$, see Fig. 63.2.

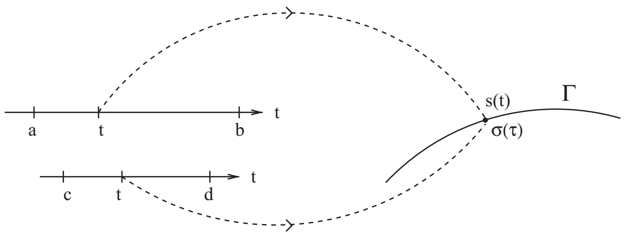


FIGURE 63.2. Reparametrization of a curve.

We now use the formula for change of integration variables and the fact that by the Chain rule

$$\sigma'(\tau) = \frac{d\sigma}{d\tau} = \frac{ds}{dt}\frac{dt}{d\tau} = s'(t)\frac{dt}{d\tau},$$

to see that, assuming $\frac{dt}{d\tau} \ge 0$,

$$\int_a^b u(s(t))\|s'(t)\|\,dt = \int_c^d u(s(t(\tau)))\|s'(t(\tau))\|\frac{dt}{d\tau}\,d\tau = \int_c^d u(\sigma(\tau))\|\sigma'(\tau)\|\,d\tau.$$

This shows that the curve integral

$$\int_\Gamma u\,ds = \int_\Gamma u\,d\sigma \tag{63.3}$$

is independent of the parametrization $s : [a, b] \to \Gamma$ or $\sigma : [c, d] \to \Gamma$ of Γ.

EXAMPLE 63.4. We reparameterize the semicircle Γ in the previous example by $s(t) = (t, \sqrt{1-t^2})$ with $-1 \le t \le 1$ and get with $u(x) = x_1^2$, integrating by parts

$$\begin{aligned}\int_\Gamma u\,ds &= \int_{-1}^{1} t\frac{t}{\sqrt{1-t^2}}\,dt = \left[-t\sqrt{1-t^2}\right]_{-1}^{1} + \int_{-1}^{1}\sqrt{1-t^2}\,dt\\ &= \int_{-\pi}^{0}\sqrt{1-\cos^2(\theta)}(-\sin(\theta))\,d\theta = \int_0^\pi \sin^2(\theta)\,d\theta = \frac{\pi}{2}.\end{aligned}$$

63.4.1 *Work and line integrals*

Let $F : \mathbb{R}^2 \to \mathbb{R}^2$ be a vector function representing a variable force, or a *force field*, defined in $\mathbb{R}^2$, and let Γ be a curve in $\mathbb{R}^2$ given by $s : [a, b] \to \mathbb{R}^2$ starting at $A = s(a)$ and ending at $B = s(b)$. Consider a particle acted upon by the force F moving along Γ from A to B, see Fig. 63.3. The projection $F_s(s(t))$ of the force $F(s(t))$ on the direction $s'(t)$ of the tangent to $s(t)$ is equal to

$$F_s(s(t)) = F(s(t)) \cdot s'(t)\frac{1}{\|s'(t)\|}. \tag{63.4}$$

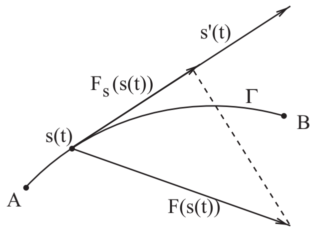


FIGURE 63.3. Force field F and curve Γ, and projection of F onto $s'(t)$.

Using the idea that "the work is equal to the projection of the force in the direction of the displacement × displacement", the *work* performed by the force $F(s(t))$ as the particle moves from $s(t_{i-1})$ to $s(t_i)$ is

$$\begin{aligned}&F(s(t_i)) \cdot s'(t_i)\frac{1}{\|s'(t_i)\|}\|s(t_i) - s(t_{i-1})\|\\ &\approx F(s(t_i)) \cdot s'(t_i)\frac{1}{\|s'(t_i)\|}\|s'(t_i)\|(t_i) - t_{i-1}) = F(s(t_i)) \cdot s'(t)(t_i - t_{i-1}).\end{aligned}$$

As above $a = t_0 < t_1 < \ldots t_{i-1} < t_i < \ldots < t_n = b$ is an increasing sequence of discrete time levels, where we think of the time steps $t_i - t_{i-1}$ as tending to zero. We are now led to define the *total work* $W(F,\Gamma)$ as the particle moves from $A = s(a)$ to $B = s(b)$ along Γ, as

$$W(F,\Gamma) = \int_a^b F(s(t)) \cdot s'(t)\, dt.$$

Setting $ds = s'(t)\, dt$, we also write

$$\int_\Gamma F \cdot ds = \int_a^b F(t) \cdot s'(t)\, dt,$$

which we call a *line integral.* To sum up, we have

$$\begin{aligned} W(F,\Gamma) &= \int_\Gamma F \cdot ds = \int_a^b F(t) \cdot s'(t)\, dt \\ &= \int_a^b (F_1(t)s_1'(t) + F_2(t)s_2'(t))\, dt. \end{aligned} \tag{63.5}$$

Alternatively, we can write

$$W(F,\Gamma) = \int_\Gamma F_s\, ds = \int_a^b F_s \|s'(t)\|\, dt = \int_\Gamma F \cdot ds, \tag{63.6}$$

with F_s being the projection of F onto $s'(t)$ according to (63.4).

EXAMPLE 63.5. Assume that $F(x) = (x_2, -x_1)$ and let Γ be given by $s(t) = (\cos(t), \sin(t))$, $0 \le t < 2\pi$. We have

$$\begin{aligned} W(F,\Gamma) &= \int_\Gamma F \cdot ds = \int_0^{2\pi} (\sin(t), -\cos(t)) \cdot (-\sin(t), \cos(t))\, dt \\ &= -\int_0^{2\pi} dt = -2\pi. \end{aligned}$$

63.5 Work and gradient fields

There is an important special case. If $F = \nabla\varphi$, that is the force field F is the *gradient field* of a *potential* $\varphi(x)$, then the Chain rule implies

$$\begin{aligned} W(F,\Gamma) &= \int_\Gamma F \cdot ds = \int_a^b \nabla\varphi(s(t)) \cdot s'(t)\, dt \\ &= \int_a^b \frac{d}{dt}\varphi(s(t))\, dt = \varphi(B) - \varphi(A). \end{aligned}$$

We conclude that if the force field F is the gradient field $F = \nabla\varphi$ of a potential $\varphi(x)$, then the work performed by F along a curve Γ from A to

B is equal to the difference $\varphi(B) - \varphi(A)$ of the values of the potential φ at the end point B and the starting point A. In other words, the work is independent of the curve from A to B. In particular, if the curve is *closed* so that $B = s(b) = s(a) = A$, then the work is zero.

Below we consider the problem of finding conditions guaranteeing that a given force $F(x)$ is the gradient of a potential so that $F(x) = \nabla\varphi(x)$ for some scalar function $\varphi(x)$.

EXAMPLE 63.6. As a basic application, we consider a *gravitational force* $F(x) = \nabla\varphi(x)$, with $\varphi(x) = 1/\|x\|$ being a *Newtonian potential*, that is

$$F(x) = -\frac{1}{\|x\|^2}\frac{x}{\|x\|}.$$

We note that $\|F(x)\| = \frac{1}{\|x\|^2}$ and that F is directed towards the origin. Thus, F corresponds to an attracting gravitational force at the origin of unit strength. We have

$$W(F,\Gamma) = \frac{1}{\|B\|} - \frac{1}{\|A\|},$$

which corresponds to the work performed as a unit mass moves in the gravitational field from a distance $\|A\|$ to the distance $\|B\|$ from the origin. In particular, if $\|A\| = \infty$, then $W(F,\Gamma) = 1/\|B\|$. We conclude that the work required to "lift" a particle of unit mass from a distance r of an attracting gravitational field of unit strength at the origin to an infinite distance is equal to $1/r$.

63.6 Using the arclength as parameter

Note that if $u(x) = 1$ for all $x \in \Gamma$ then

$$\int_\Gamma ds = \int_\Gamma 1\, ds = \int_\Gamma u(x)\, ds(x) = \int_a^b \|s'(t)\|\, dt \tag{63.7}$$

is the length of the curve $\Gamma = s(I)$ with $I = [a,b]$. In particular,

$$\sigma(\bar{t}) = \int_a^{\bar{t}} \|s'(t)\|\, dt$$

is the *arclength* of the part of the curve from $s(a)$ to $s(\bar{t})$. The Fundamental Theorem of Calculus implies

$$\sigma'(\bar{t}) = \|s'(\bar{t})\|. \tag{63.8}$$

We may now choose the arclength $\sigma = \sigma(t)$ as the parameter instead of t since to each t, there is a unique arclength $\sigma(t)$ and vice versa. This gives a reparameterization of $s(t) = \bar{s}(\sigma)$ with

$$\|\bar{s}'(\sigma)\| = \|\frac{ds}{dt}\||\frac{dt}{d\sigma}| = \|s'(t)\|\frac{1}{|\sigma'(t)|} = \frac{\|s'(t)\|}{\|s'(t)\|} = 1.$$

We conclude that if the arclength σ to used to parameterize the curve $s : I \to \mathbb{R}^2$, then $\|s'(\sigma)\| = 1$ and, see Fig. 63.4,

$$L(\Gamma) = \int_\Gamma ds = \int_0^{L(\Gamma)} d\sigma.$$

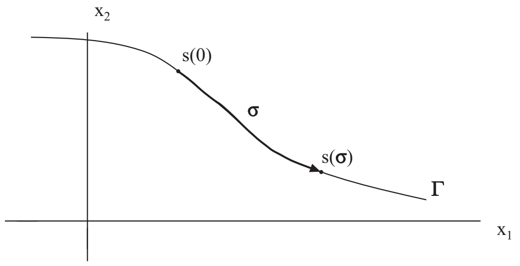


FIGURE 63.4. A curve Γ parameterized by arclength σ.

63.6.1 The curvature of a plane curve

The *curvature* of a curve $s : [a, b] \to \mathbb{R}^2$ measures of how quickly the curve bends as we move along the curve. It is defined by

$$\kappa = \frac{d\theta}{d\sigma},$$

where θ is the polar angle of the tangent vector $s' = (s_1', s_2')$ defined by $\theta(t)) = \tan^{-1}(s_2'/s_1')$ and σ is arclength. In the case of a straight line, the polar angle $\theta(t)$ is constant and the curvature is zero, see Fig. 63.5.

The arc length $\sigma(t)$ satisfies, recalling (63.8), $\frac{d\sigma}{dt} = |s'|$, and thus $\frac{dt}{d\sigma} = |s'|^{-1}$. The chain rule implies

$$\kappa(t) = \frac{d\theta}{dt}\frac{dt}{d\sigma} = \frac{\theta'(t)}{\|s'(t)\|}.$$

Computing $\theta'(t)$, we find that

$$\kappa(t) = \frac{s_1'(t)s_2''(t) - s_1''(t)s_2'(t)}{\left(s_1'(t)^2 + s_2'(t)^2\right)^{3/2}}.$$

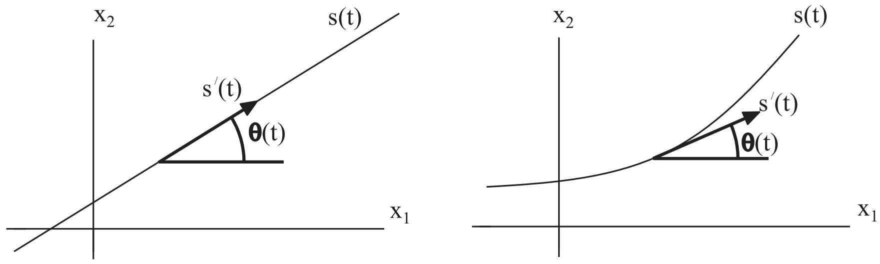


FIGURE 63.5. The polar angle θ of the tangent vector of a straight line is constant as shown on the right. The tangent vector of a curve that bends, like the example on the left, has a different polar angle at each point.

In particular if the curve is parameterized by $s(x_1) = (x_1, f(x_1))$, where $f : \mathbb{R} \to \mathbb{R}$ has two continuous derivatives, then the curvature at the point $(x_1, f(x_1))$, is given by

$$\kappa(x_1) = \frac{f''(x_1)}{\big(1 + (f'(x_1))^2\big)^{3/2}}.$$

We define the *circle of curvature* at a point $P = s(t)$ on a curve $s : [a, b] \to \mathbb{R}^2$, as the circle of radius $|\kappa|^{-1}(t)$ (assuming $\kappa \neq 0$) that shares the same tangent line as Γ at P and points to the left of T if $\kappa > 0$ and to the right if $\kappa < 0$, see Fig. 63.6. The *radius of curvature* at P is $|\kappa|^{-1}(t)$.

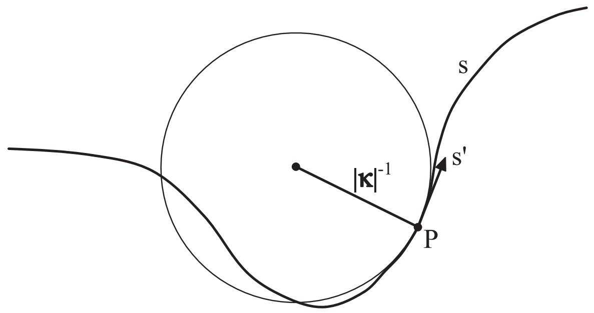


FIGURE 63.6. The circle of curvature of Γ at P.

63.7 Extension to curves in $\mathbb{R}^n$

The definitions of integrals over curves and line integrals directly extend to curves in $\mathbb{R}^n$ represented by $s : [a, b] \to \mathbb{R}^n$ with $n \geq 2$.

EXAMPLE 63.7. Consider the circular helix Γ in $\mathbb{R}^3$ given by $s(t) = (\cos(t), \sin(t), t)$, $0 \leq t \leq 20\pi$, and let $u(x) = x_3^2$. We have since $s'(t) = (-\sin(t), \cos(t), 1)$ and thus $\|s'(t)\| = \sqrt{2}$,

$$\int_\Gamma u\, ds = \int_0^{20\pi} t^2\sqrt{2}\, dt = \frac{\sqrt{2}}{3}(20\pi)^3.$$

Chapter 63 Problems

63.1. (a) Compute the length of (a) the *catenary* (hanging chain curve) given by $s(t) = (t, \cosh(t))$ with $-1 \leq t \leq 1$, (b) the *circular helix* $s(t) = (\cos(t), \sin(t), t)$ with $0 \leq t \leq 4\pi$, (c) the *cycloid* $s(t) = (t - \sin(t), 1 - \cos(t))$ with $0 \leq t \leq 2\pi$, (d) the *semi-cubical parabola* $s(t) = (t^3, t^2)$ with $0 \leq t \leq 2$, (e) the *four-cusped hypocycloid or astroid* $s(t) = (\cos^3(t), \sin^3(t))$.

63.2. Let Γ be the circular helix $s(t) = (\cos(t), \sin(t), t)$ with $t \in [0, 2\pi)$. Compute the value of the curve integral $\int_\Gamma u\, ds$ for (a) $u(x) = 1$, (b) $u(x) = x_3$, (c) $u(x) = x_1x_2x_3$.

63.3. Compute the curve integral $\int_\Gamma x_1x_2\, ds$, where (a) Γ is the part of the unit circle in the x_1x_2-plane from $(1,0,0)$ to $(0,1,0)$, (b) Γ is the part of the unit square in the x_1x_2-plane from $(1,0,0)$ to $(0,1,0)$. (c) Γ is the shortest path from $(1,0,0)$ to $(0,1,0)$.

63.4. (a) Compute the line integral $\int_\Gamma x \cdot ds$ where Γ is the unit circle in the x_1x_2-plane. (b) Try other choices of closed curves Γ and evaluate the integral.

63.5. Compute the line integral $\int_\Gamma F \cdot ds$ with Γ the unit circle in the x_1x_2 plane and (a) $F(x) = \frac{(x_1,x_2)}{|x|^2}$, (b) $F(x) = \frac{(-x_2,x_1)}{|x|^2}$. Does the result depend on whether you integrate around the unit circle clockwise or counter-clockwise?

63.6. A particle is moved counter-clockwise around the square $0 \leq x_1, x_2 \leq 1$, $x_3 = 0$ under the action of the force field $f(x) = ((x_1 - x_2)^2, 2x_2 + x_1^2, x_1)$. Compute the work done.

63.7. Let $f(x) = (2x_1 + x_2, 3x_1 - 2x_2)$. Compute $\int_\Gamma f \cdot ds$ with Γ given by (a) the straight line from $(0,0)$ to $(1,1)$, (b) the parabola $x_2 = x_1^2$ from $(0,0)$ to $(1,1)$, (c) the curve $x_2 = sin(\pi x_1/2)$ from $(0,0)$ to $(1,1)$, (d) the curve $x_2 = x_1^n$ with $n > 0$ from $(0,0)$ to $(1,1)$.

63.8. Compute the integral of $u = x_1\, x_2$ over the boundary of the unit square $[0,1] \times [0,1]$.

63.9. Find the circle of curvature of $x_2 = x_1^2$ at $x_1 = 0$.

63.10. Find the curvature of the plane curve $\big(R\cos(\theta), R\sin(\theta)\big)$ where R is constant. Conclude that the curvature of a circle of radius R is R^{-1}.

63.11. Verify the two formulas for the curvature.

63.12. (a) Compute the curvature of the curve (x_1, x_1^2). (b) Do the same for (x_1, x_1^3), and then discuss what happens at the inflection point.

63.13. Consider a hanging chain described by a function $y(x)$ with $-1 \leq x \leq 1$ and $y(-1) = y(1)$. Let for $0 \leq x \leq 1$, $T(x)$ be the modulus of the chain force at x, and let $s(x)$ be the length of the chain from 0 to x. Derive the vertical equilibrium equation

$$y'(x) = cs(x) = c\int_0^1 \sqrt{1 + (y'(x))^2}\, dx,$$

with c a constant. Show that this equation is satisfied with $y'(x) = \sinh(\frac{x}{c})$, and conclude that $y(x) = c\cosh(\frac{x}{c})$.

63.14. Find the direction of the tangent at the point $(1, 1, 1)$ of the curve cut out on the surface $x_1^2 + x_1^2 x_2 + x_2^2 x_3 + x_3^2 = 0$. Hint: Use implicit differentiation.

63.15. Show that if a plane curve Γ is represented in polar coordinates $(\rho(\theta), \theta)$ with $\rho(theta)$ a function of θ and $a \leq \theta \leq b$, then $ds^2 = \rho^2\, d\theta^2 + d\rho^2$ and thus

$$L(\Gamma) = \int_a^b (\rho^2 + (\rho')^2)^{1/2}\, d\theta.$$

Compute the the length of the *cardioid* $\rho = (1 - \cos(\theta)$ with $0 \leq \theta \leq 2\pi$.

63.16. Compute the length of a string which is wound around a circular cylinder with a uniform pitch.

64 Double integrals

To understand this for sense it is not required that a man should be a geometrician or a logician, but that he should be mad. ["This" is that the volume generated by revolving the region under 1/x from 1 to infinity has finite volume.] (Hobbes 1588-1679)

He was 40 years old before he looked on geometry; which happened accidentally. Being in a gentleman's library, Euclid's Elements lay open, and "twas the 47 El. libri I" [Pythagoras' Theorem]. He read the proposition . "By God", sayd he, "this is impossible:" So he reads the demonstration of it, which referred him back to such a proposition; which proposition he read. That referred him back to another, which he also read. Et sic deinceps, that at last he was demonstratively convinced of that trueth. This made him in love with geometry. (About Thomas Hobbes by John Aubrey 1626-1697)

64.1 Introduction

We have studied the integral

$$\int_0^1 f(x)\,dx,$$

where $f : [0,1] \to \mathbb{R}$ is a Lipschitz continuous function of one variable. We call this a *one-dimensional integral*. We generalize this idea to the *double integral*

$$\int_0^1 \int_0^1 f(x_1, x_2)\,dx_1\,dx_2, \tag{64.1}$$

which has *two integration variables* x_1 and x_2 that run from 0 to 1. Here $f : Q \to \mathbb{R}$ is a Lipschitz continuous function defined on the unit square $Q = [0,1] \times [0,1] = \{x = (x_1, x_2) : 0 \le x_1 \le 1, 0 \le x_2 \le 1\}$, satisfying

$$|f(x) - f(y)| \le L_f \|x - y\| \quad \text{for } x, y \in Q. \tag{64.2}$$

64.2 Double integrals over the unit square

Recall that we define the one dimensional integral as

$$\int_0^1 f(x)\, dx = \lim_{n\to\infty} \sum_{i=1}^{N} f(x_i^n) h_n, \tag{64.3}$$

where $0 = x_0^n < x_1^n < ... < x_N^n = 1$ is a subdivision of the interval $[0,1]$ with $x_i^n = ih_n$, $i = 1, ..., N$ and $h_n = 2^{-n}$ and $N = 2^n$.

To define the double integral, we let $0 = x_{1,0}^n < x_{1,1}^n < ... < x_{1,N}^n = 1$ and $0 = x_{2,0}^n < x_{2,1}^n < ... < x_{2,N}^n = 1$ be a subdivisions of the interval $[0,1]$ with $x_{1,i}^n = ih_n$, $i = 0, ..., N$, and $x_{2,j}^n = jh_n$, $j = 0, ..., N$, where $h_n = 2^{-n}$ and $N = 2^n$. This corresponds to a subdivision of the unit square $Q = [0,1]\times[0,1]$ into sub-squares $Q_{i,j}^n = I_i^n \times J_j^n$ of area $h_n h_n$, where $I_i^n = (x_{1,i-1}^n, x_{1,i}^n]$ $J_j^n = (x_{2,j-1}^n, x_{2,j}^n]$, where $i, j = 1, ..., N$, see Fig. 64.1.

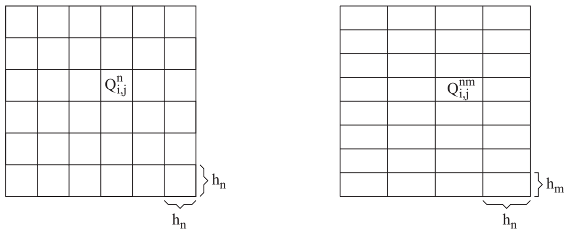


FIGURE 64.1. Partition of the unit square Q into quadratic or rectangular sub-domains Q_{ij}^n or Q_{ij}^{nm}.

We shall prove that the limit $\lim_{n\to\infty} S_n$ exists, where

$$S_n = \sum_{i=1}^{N} \sum_{j=1}^{N} f(x_{1,i}^n, x_{2,j}^n) h_n h_n \tag{64.4}$$

is a *Riemann sum* over all the sub-squares $Q_{i,j}^n$. We define

$$\int_0^1 \int_0^1 f(x_1, x_2)\, dx_1\, dx_2 = \lim_{n\to\infty} \sum_{i=1}^{N} \sum_{j=1}^{N} f(x_{1,i}^n, x_{2,j}^n) h_n h_n. \tag{64.5}$$

We begin by estimating the difference $S_n - S_{n+1}$ with the goal of proving that $\{S_n\}$ is a Cauchy sequence. Each sub-square $Q^n_{i,j}$ consists of the four sub-squares $Q^{n+1}_{2i,2j}$, $Q^{n+1}_{2i-1,2j}$ $Q^{n+1}_{2i,2j-1}$, and $Q^{n+1}_{2i-1,2j-1}$, see Fig. 64.2. We have

$$S_n - S_{n+1} = \sum_{i=1}^{N}\sum_{j=1}^{N} a_{ij} h_n h_n,$$

where, see Fig. 64.2,

$$a_{ij} = f(x^n_{1,i}, x^n_{2,j}) - \frac{1}{4}\big(f(x^{n+1}_{1,2i}, x^{n+1}_{2,2j}) + f(x^{n+1}_{1,2i-1}, x^{n+1}_{2,2j}) \\ + f(x^{n+1}_{1,2i}, x^{n+1}_{2,2j-1}) + f(x^{n+1}_{1,2i-1}, x^{n+1}_{2,2j-1})\big).$$

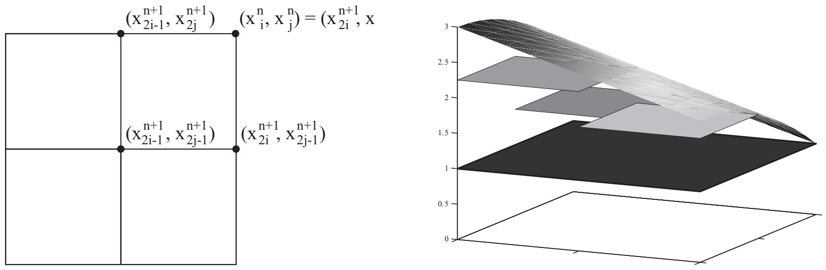


FIGURE 64.2. On the left: Q^n_{ij} and four sub-squares and quadrature point. On the right: A function $f(x_1, x_2)$ and its piecewise constant approximation on Q^n_{ij} and on the four sub-squares.

The Lipschitz continuity condition (64.2) implies

$$|a_{ij}| \le \frac{1}{4} L_f(h_{n+1} + h_{n+1} + \sqrt{2}h_{n+1}) \le L_f h_{n+1},$$

and thus

$$|S_n - S_{n+1}| \le \sum_{i=1}^{N}\sum_{j=1}^{N} |a_{ij}| h_n h_n \le L_f h_{n+1} \sum_{i=1}^{N}\sum_{j=1}^{N} h_n h_n = L_f h_{n+1}.$$

The usual arguments show that for $m > n$,

$$|S_n - S_m| \le 2L_f h_{n+1} = L_f h_n,$$

which proves that $\{S_n\}$ is a Cauchy sequence and thus converges to a real number. We decide, following our dear friends Leibniz and Cauchy as usual, to denote this real number by

$$\int_0^1 \int_0^1 f(x_1, x_2)\, dx_1\, dx_2 = \lim_{n\to\infty} S_n = \lim_{n\to\infty} \sum_{i=1}^{N}\sum_{j=1}^{N} f(x^n_{1,i}, x^n_{2,j}) h_n h_n.$$

We shall also use the notation

$$\int_Q f(x)\,dx = \int_0^1 \int_0^1 f(x_1, x_2)\,dx_1\,dx_2.$$

We summarize as follows:

Theorem 64.1 *If $f : [0,1] \times [0,1] \to \mathbb{R}$ is Lipschitz continuous, then the limit*

$$\lim_{n\to\infty} \sum_{i=1}^{N} \sum_{j=1}^{N} f(x_{1,i}^n, x_{2,j}^n) h_n h_n,$$

exists, where $h_n = 2^{-n}$ and $N = 2^n$, $x_{1,i}^n = ih_n$, $x_{2,j}^n = jh_n$, and we define

$$\int_Q f(x)\,dx = \int_0^1 \int_0^1 f(x_1, x_2)\,dx_1\,dx_2 = \lim_{n\to\infty} \sum_{i=1}^{N} \sum_{j=1}^{N} f(x_{1,i}^n, x_{2,j}^n) h_n h_n. \tag{64.6}$$

In general, the partitions in x_1 and x_2 can be independent, leading to Riemann sums of the form

$$S_{nm} = \sum_{i=1}^{N} \sum_{j=1}^{M} f(x_{1,i}^n, x_{2,j}^m) h_n h_m, \tag{64.7}$$

where $h_n = 2^{-n}$ and $N = 2^n$, $h_m = 2^{-m}$ and $M = 2^m$. This corresponds to a subdivision of Q into sub-squares $Q_{ij}^{nm} = I_i^n \times J_j^m$. The proof above directly generalizes to prove that if $\overline{n} \geq n$ and $\overline{m} \geq m$ then

$$|S_{nm} - S_{\overline{nm}}| \leq L_f \max(h_n, h_m).$$

This proves the following generalization of the previous theorem.

Theorem 64.2 *Suppose $f : [0,1] \times [0,1] \to \mathbb{R}$ is Lipschitz continuous. Then the following limit exists*

$$\lim_{n,m\to\infty} \sum_{i=1}^{N} \sum_{j=1}^{M} f(x_{1,i}^n, x_{2,j}^m) h_n h_m,$$

where $h_n = 2^{-n}$, $N = 2^n$, $h_m = 2^{-m}$, $M = 2^m$, $x_{1,i}^n = ih_n$, $x_{2,j}^m = jh_m$, and

$$\int_Q f(x)\,dx = \int_0^1 \int_0^1 f(x_1, x_2)\,dx_1\,dx_2 = \lim_{n,m\to\infty} \sum_{i=1}^{N} \sum_{j=1}^{M} f(x_{1,i}^n, x_{2,j}^m) h_n h_m. \tag{64.8}$$

64.3 The double integral over Q through repeated one-dimensional integration

To compute the Riemann sum S_{nm}, we have to perform a summation over all the sub-squares Q_{ij}^{nm} covering Q. The summation may be performed in different orders, row by row, column by column, or in some other order. We thus obtain the following alternative expressions for the double integral of $f(x_1, x_2)$ over Q:

$$\begin{aligned}\int_0^1 \int_0^1 f(x_1, x_2)\, dx_1\, dx_2 &= \lim_{n,m\to\infty} \sum_{i=1}^{N} \sum_{j=1}^{M} f(x_{1,i}^n, x_{2,j}^m) h_n h_m \\ &= \lim_{n,m\to\infty} \sum_{i=1}^{N} \big(\sum_{j=1}^{M} f(x_{1,i}^n, x_{2,j}^m) h_m\big) h_n \\ &= \lim_{n,m\to\infty} \sum_{j=1}^{M} \big(\sum_{i=1}^{N} f(x_{1,i}^n, x_{2,j}^m) h_n\big) h_m,\end{aligned}$$

where $\sum_{i=1}^N \sum_{j=1}^M$ indicates an arbitrary order of summation, $\sum_{i=1}^N \big(\sum_{j=1}^M\big)$ summation column by column, and $\sum_{j=1}^M \big(\sum_{i=1}^N\big)$ summation row by row over the subdomains Q_{ij}^{nm} of Q in the x_1x_2-plane, see Fig. 64.3

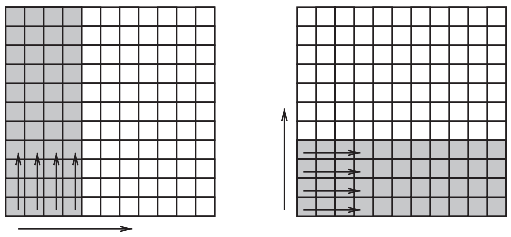


FIGURE 64.3. Different orders of summation

We can also perform the limits with respect to n and m independently, and we then arrive at the formula

$$\begin{aligned}\int_0^1 \int_0^1 f(x_1, x_2)\, dx_1\, dx_2 &= \lim_{n,m\to\infty} \sum_{i=1}^{N} \sum_{j=1}^{M} f(x_{1,i}^n, x_{2,j}^m) h_n h_m \\ &= \lim_{n\to\infty} \sum_{i=1}^{N} \big(\lim_{m\to\infty} \sum_{j=1}^{M} f(x_{1,i}^n, x_{2,j}^m h_m\big) h_n \\ &= \lim_{m\to\infty} \sum_{j=1}^{M} \big(\lim_{n\to\infty} \sum_{i=1}^{N} f(x_{1,i}^n, x_{2,j}^m h_n\big) h_m.\end{aligned}$$

This corresponds to the following formula:

$$\int_0^1 \int_0^1 f(x_1, x_2)\, dx_1\, dx_2 = \int_0^1 \big(\int_0^1 f(x_1, x_2)\, dx_2 \big)\, dx_1$$
$$= \int_0^1 \big(\int_0^1 f(x_1, x_2)\, dx_1 \big)\, dx_2,$$

or

$$\int_0^1 \int_0^1 f(x_1, x_2)\, dx_1\, dx_2 = \int_0^1 g_1(x_1)\, dx_1 = \int_0^1 g_2(x_2)\, dx_2,$$

where

$$g_1(x_1) = \int_0^1 f(x_1, x_2)\, dx_2 = \lim_{m\to\infty} \sum_{j=1}^{M} f(x_1, x_{2,j}^m) h_m$$

and

$$g_2(x_2) = \int_0^1 f(x_1, x_2)\, dx_1 = \lim_{n\to\infty} \sum_{i=1}^{N} f(x_{1,i}^n, x_2) h_n$$

define functions $g_1(x_1)$ and $g_2(x_2)$ of x_1 and x_2 respectively. In other words, the double integral of $f(x_1, x_2)$ over $[0,1]\times[0,1]$ equals the integral of $g_2(x_2)$ over $[0,1]$,

$$\int_0^1 \int_0^1 f(x_1, x_2)\, dx_1\, dx_2 = \int_0^1 g_2(x_2) dx_2 = \lim_{n\to\infty} \sum_{j=1}^{M} g_2(x_{2,j}^m) h_m,$$

and equals the integral of $g_1(x_1)$ over $[0,1]$,

$$\int_0^1 \int_0^1 f(x_1, x_2)\, dx_1\, dx_2 = \int_0^1 g_1(x_1) dx_1 = \lim_{n\to\infty} \sum_{i=1}^{N} g_1(x_{1,i}^n) h_n.$$

We conclude that a double integral can be computed by repeated, or iterated, integration in one dimension. We may summarize this experience as follows:

Theorem 64.3 *If $f : [0,1] \times [0,1] \to \mathbb{R}$ is Lipschitz continuous, then*

$$\int_Q f(x)\, dx = \int_0^1 \int_0^1 f(x_1, x_2)\, dx_1\, dx_2 =$$
$$= \int_0^1 \big(\int_0^1 f(x_1, x_2)\, dx_2 \big)\, dx_1 = \int_0^1 \big(\int_0^1 f(x_1, x_2)\, dx_1 \big)\, dx_2.$$

We can interpret the statement of this theorem as a *change of order of integration* in the sense that integrating with respect to x_1 and then with respect to x_2 gives the same result as integrating first with respect to x_2 and then with respect to x_1. The usual way to evaluate a double integral is to use iterated one-dimensional integration in some order.

EXAMPLE 64.1.

With $Q = [0,1] \times [0,1]$,

$$\int_Q x_1 x_2^3\, dx = \int_0^1 \int_0^1 x_1 x_2^3\, dx_1 dx_2 = \int_0^1 x_1 \big(\int_0^1 x_2^3\, dx_2 \big) dx_1$$
$$= \int_0^1 x_1 \big[\frac{x_2^4}{4} \big]_0^1 dx_1 = \frac{1}{4} \int_0^1 x_1 dx_1 = \frac{1}{4} \big[\frac{x_1^2}{2} \big]_0^1 = \frac{1}{8}.$$

$$\int_Q x_1 x_2^3\, dx = \int_0^1 \int_0^1 x_1 x_2^3\, dx_1 dx_2 = \int_0^1 x_2^3 \big(\int_0^1 x_1\, dx_1 \big) dx_2$$
$$= \int_0^1 x_2^3 \big[\frac{x_1^2}{2} \big]_0^1 dx_2 = \frac{1}{2} \int_0^1 x_2^3 dx_2 = \frac{1}{2} \big[\frac{x_2^4}{4} \big]_0^1 = \frac{1}{8}.$$

Alternatively, we may first integrate with respect to x_1 and then with respect to x_2 to get,

$$\int_Q x_1 x_2^3\, dx = \int_0^1 \int_0^1 x_1 x_2^3\, dx_1 dx_2 = \int_0^1 x_2^3 \big(\int_0^1 x_1\, dx_1 \big) dx_2$$
$$= \int_0^1 x_2^3 \big[\frac{x_1^2}{2} \big]_0^1 dx_2 = \frac{1}{2} \int_0^1 x_2^3 dx_2 = \frac{1}{2} \big[\frac{x_2^4}{4} \big]_0^1 = \frac{1}{8}.$$

64.4 Generalization to an arbitrary rectangle

The double integral defined on the unit square generalizes directly to integrals over arbitrary rectangles $Q = [a_1, b_1] \times [a_2, b_2]$ with sides parallel to the axis. If $f : Q \to \mathbb{R}$ is Lipschitz continuous, then

$$\int_Q f(x)\, dx = \int_Q f(x_1, x_2)\, dx_1 dx_2 = \int_{a_1}^{b_1} \big(\int_{a_2}^{b_2} f(x_1, x_2)\, dx_2 \big)\, dx_1$$
$$= \int_{a_2}^{b_2} \big(\int_{a_1}^{b_1} f(x_1, x_2)\, dx_1 \big)\, dx_2.$$

64.5 Interpreting the double integral as a volume

The sum

$$\sum_{i=1}^{N} \sum_{j=1}^{N} f(x_{1,i}^n, x_{2,j}^n) h_n h_n \tag{64.9}$$

represents the sum of the volumes

$$f(x_{1,i}^n, x_{2,j}^n) h_n h_n \tag{64.10}$$

of thin boxes with cross-section of area $h_n h_n$ and height $f(x_{1,i}^n, x_{2,j}^n)$. Intuitively, this is an approximation of the volume under the graph of $f(x_1, x_2)$ with (x_1, x_2) varying over Q. It is thus natural to define the volume $V(f, Q)$ under the graph of $f(x_1, x_2)$ over Q to be

$$V(f,Q) = \int_0^1 \int_0^1 f(x_1, x_2)\, dx_1\, dx_2 \tag{64.11}$$

EXAMPLE 64.2. We compute the volume of a pyramid of height 1 with base $[0,2] \times [0,2]$, see Fig. 64.4. One quarter of the volume is equal to the integral $\int_Q f(x)\, dx$, where $Q = [0,1] \times [0,1]$, $f(x) = x_2$ for $x \in Q$ such that $x_2 \le x_1$ and $f(x) = x_1$ for $x \in Q$ such that $x_1 \le x_2$. We have

$$\begin{aligned} V(f,Q) &= \int_Q f(x)\, dx = \int_0^1 \Big(\int_0^{x_1} x_2\, dx_2 + \int_{x_1}^1 x_1\, dx_2\Big)\, dx_1 \\ &= \int_0^1 \Big(\frac{x_1^2}{2} + x_1(1 - x_1)\, dx_1 = \Big[\frac{x_1^2}{2} - \frac{x_1^3}{6}\Big]_0^1 = \frac{1}{2} - \frac{1}{6} = \frac{1}{3}. \end{aligned}$$

We conclude that the volume of the pyramid is equal to $\frac{4}{3}$. This agrees with the standard formula stating that the volume of a pyramid is equal to $\frac{1}{3}Bh$, where B is the area of the base and h is the height.

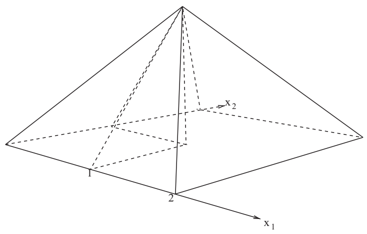


FIGURE 64.4. Volume of pyramid

64.6 Extension to general domains

We next define the double integral of a function $f(x)$ over a more general domain Ω in the plane. We start by assuming that the boundary Γ of Ω is described by two curves $x_2 = \gamma_1(x_1)$ and $x_2 = \gamma_2(x_1)$ for $0 \le x_1 \le 1$, as shown in Fig. 64.5, so that $\Omega = \{x \in [0,1] \times \mathbb{R} : \gamma_1(x_1) \le x_2 \le \gamma_2(x_1)\}$ We assume that the functions $\gamma_i : [0,1] \to \mathbb{R}$ are Lipschitz continuous with Lipschitz constant L_γ. We further assume that $f : \Omega \to \mathbb{R}$ is Lipschitz continuous with Lipschitz constant L_f.

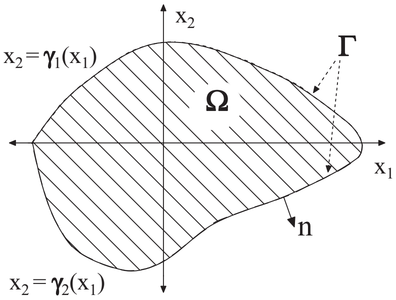


FIGURE 64.5. The domain Ω in the plane

We assume that Ω is contained in the unit square Q. We partition Q as above into squares $I_i^n \times J_j^n$ of area $h_n h_n$, where $I_i^n = (x_{1,i-1}^n, x_{1,i}^n]$ $J_j^n = (x_{2,j-1}^n, x_{2,j}^n]$. We denote by ω_n the set of indices (i,j) such that the square $I_i^n \times J_j^n$ intersects Ω, and we let Ω_n be the union of the squares $I_i^n \times J_j^n$ with indices $(i,j) \in \omega_n$. In other words, Ω_n is an approximation of Ω consisting of all the squares $I_i^n \times J_j^n$ in Q that intersect Ω. We consider the Riemann sum

$$S_n = \sum_{(i,j)\in\omega_n} f(x_{1,i}^n, x_{2,j}^n) h_n h_n. \tag{64.12}$$

We shall prove that $\lim_{n\to\infty} S_n$ exists and then naturally define

$$\int_\Omega f(x)\,dx = \lim_{n\to\infty} \sum_{(i,j)\in\omega_n} f(x_{1,i}^n, x_{2,j}^n) h_n h_n. \tag{64.13}$$

To this end, we estimate the difference $S_n - S_{n+1}$, which now has contributions from two sources; (i) from the variation of $f(x)$ over each sub-square $I_i^n \times J_j^n$, and (ii) from the difference between Ω_n and Ω_{n+1}.

The first contribution can be shown to be bounded by $L_f h_n$ by arguing just as for integration over a square. The second contribution is bounded by $2A(1+L_\gamma)h_n$, where A is a bound for $|f(x)|$, that is $|f(x)| \leq A$ for $x \in \Omega$. This follows from the observation that if a square $I_i^n \times J_j^n$ of Ω_n, is entirely outside or inside Ω, then so are all the four squares of Ω_{n+1} within $I_i^n \times J_j^n$. The difference between Ω_n and Ω_{n+1} arises from the squares $I_i^n \times J_j^n$ which are partly inside and partly outside Ω. The area of these squares is bounded by $2L_\gamma h_n$, where the factor 2 arises from the fact that there are two curves γ_1 and γ_2, see Fig. 64.6. The difference in area between Ω_n and Ω_{n+1} is thus bounded by $2L_\gamma h_n$.

Together, this shows that

$$|S_n - S_{n+1}| \leq (L_f + 2AL_\gamma)h_n, \tag{64.14}$$

which as proves that $\lim_{n\to\infty} S_n$ exists. We summarize as follows:

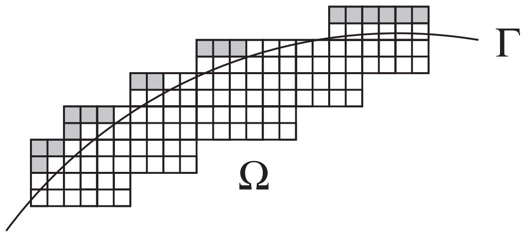


FIGURE 64.6. Approximation of integral over general domain

Theorem 64.4 *Let* $\Omega = \{x \in [0,1] \times \mathbb{R} : \gamma_2(x_1) \leq x_2 \leq \gamma_1(x_1)\}$, *where* $\gamma_i : [0,1] \to \mathbb{R}$ *are Lipschitz continuous, and let* $f : \Omega \to \mathbb{R}$ *be Lipschitz continuous. Then* $\lim_{n\to\infty} S_n$ *exists, where* S_n *is the Riemann sum defined by* (64.12), *and we define*

$$\int_\Omega f(x)\, dx = \int_\Omega f(x_1, x_2)\, dx_1 dx_2 = \lim_{n\to\infty} S_n. \qquad (64.15)$$

64.7 Iterated integrals over general domains

The integral of a function $f(x)$ over a domain $\Omega = \{x \in [0,1] \times \mathbb{R} : \gamma_2(x_1) \leq x_2 \leq \gamma_1(x_1)\}$ may be computed by iterated integration in one dimension as follows

$$\int_\Omega f(x)\, dx = \int_\Omega f(x_1, x_2)\, dx_1\, dx_2 = \int_0^1 \left(\int_{\gamma_2(x_1)}^{\gamma_1(x_1)} f(x_1, x_2)\, dx_2 \right) dx_1. \qquad (64.16)$$

This is another way of expressing the fact that,

$$\begin{aligned} \int_\Omega f(x)\, dx &= \lim_{n\to\infty} \sum_{(i,j)\in\omega_n} f(x_{1,i}^n, x_{2,j}^n) h_n h_n \\ &= \lim_{n\to\infty} \sum_{i=1}^N \Big(\sum_{j:(i,j)\in\omega_n} f(x_{1,i}^n, x_{2,j}^n) h_n \Big) h_n \end{aligned}$$

The role of x_1 and x_2 may be interchanged and the integral is independent of the particular representation of Γ. To handle a more general domain Ω, we split Ω into appropriate sub-domains Ω_j and define $\int_\Omega f\, dx = \sum_j \int_{\Omega_j} f\, dx$. Again the integral of f over Ω represents the volume of the domain under the graph of f over Ω.

Evaluation of an integral over a two-dimensional domain by repeated integration was used by Euler in 1738, when he computed the gravitational attraction of an elliptic lamina.

EXAMPLE 64.3. We compute the double integral

$$I = \int_\Omega (x_1^2 + x_2)\,dx,$$

over the domain $\Omega = \{x \in \mathbb{R}^2 : x_1^2 \le x_2 \le x_1,\, 0 \le x_1 \le 1\}$. We have

$$\begin{aligned} I &= \int_0^1 \big(\int_{x_1^2}^{x_1} (x_1^2 + x_2)\,dx_2\big)dx_1 = \int_0^1 \big[x_1^2 x_2 + \frac{x_2^2}{2}\big]_{x_1^2}^{x_1} dx_1 \\ &= \int_0^1 (x_1^3 + \frac{x_1^2}{2} - x_1^4 - \frac{x_1^4}{2})dx_1 = \frac{1}{4} + \frac{1}{6} - \frac{1}{5} - \frac{1}{10} = \frac{7}{60}. \end{aligned}$$

EXAMPLE 64.4. We compute the double integral

$$I = \int_\Omega \frac{1}{x_2}\,dx$$

over the domain $\Omega = \{x \in \mathbb{R}^2 : 1 \le x_2 \le \exp(x_1),\, 0 \le x_1 \le 1\}$. We have

$$I = \int_0^1 \big(\int_1^{\exp(x_1)} \frac{1}{x_2}\,dx_2\big)dx_1 = \int_0^1 \big[\log(x_2)\big]_1^{\exp(x_1)} dx_1 = \int_0^1 x_1 dx_1 = \frac{1}{2}.$$

64.8 The area of a two-dimensional domain

We define the *area* $A(\Omega)$ of a domain Ω in $\mathbb{R}^2$ by

$$A(\Omega) = \int_\Omega dx, \tag{64.17}$$

i.e. by integration of the constant function $f(x) = 1$ over Ω. If $\Omega = \{x \in [0,1] \times \mathbb{R} : \gamma_1(x_1) \le x_2 \le \gamma_2(x_1)\}$, then

$$A(\Omega) = \int_0^1 \left(\int_{\gamma_2(x_1)}^{\gamma_1(x_1)} dx_2\right) dx_1 = \int_0^1 \big(\gamma_2(x_1) - \gamma_1(x_1)\big)\,dx_1,$$

which conforms with the previous formula of the area between the curves $\gamma_1(x_1)$ and $\gamma_1(x_1)$ as the integral of the difference $\gamma_2(x_1) - \gamma_1(x_1)$.

EXAMPLE 64.5. The area of the triangle Ω with corners at $(0,0)$, $(1,0)$ and $(1,1)$, can be computed as follows

$$A(\Omega) = \int_\Omega dx = \int_0^1 \big(\int_0^{x_1} dx_2\big)\,dx_1 = \int_0^1 \frac{1}{2}\,dx_1 = \frac{1}{2}.$$

64.9 The integral as the limit of a general Riemann sum

We defined the integral using uniform subdivisions in x_1 and x_2, resulting in approximate subdivisions of a given domain Ω in $\mathbb{R}^2$ into squares or rectangles. We can however use more general subdivisions of Ω. Suppose that $f:\Omega\rightarrow\mathbb{R}$ is a Lipschitz continuous function and that the boundary of a domain Ω can be made up of pieces of Lipschitz curves $x_2=\gamma(x_1)$ or $x_1=\gamma(x_2)$. For $N=1,2,...,$ we divide Ω into a collection $\{\Omega_i\}_{i=1}^N$ of pairwise disjoint sets Ω_i such that the union of the Ω_i is equal to Ω. Let $d\Omega_i$ be the area of Ω_i and let d_N be the maximal diameter of Ω_i for $i=1,...,N$, see Fig. 64.7. We assume that d_N tends to zero as N tends to infinity.

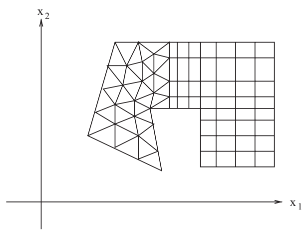


FIGURE 64.7. Subdivison of general domain

The arguments used above show that

$$\int_\Omega f(x)\,dx=\lim_{N\rightarrow\infty}\sum_{i=1}^N f(x_i)d\Omega_i, \tag{64.18}$$

where x_i is a point in Ω_i for $i=1,..,N$. The first step in proving this result is to use the estimate

$$|f(x)-f(y)|\leq L_f d_N \quad \text{if } x,y\in\Omega_i, \tag{64.19}$$

which implies that the variation of $f(x)$ with x ranging over Ω_i is small if the diameter of Ω_i is small. The second step involves the Lipschitz continuity of the boundary of Ω and the boundedness of $f(x)$. By the way, a byproduct of the proof of this result is the estimate

$$|\int_\Omega f(x)\,dx-\sum_{i=1}^N f(x_i)d\Omega_i|\leq L_f d_N A(\Omega), \tag{64.20}$$

where $A(\Omega)$ is the area of Ω.

64.10 Change of variables in a double integral

We next extend the idea of changing variables in a one-dimensional integral to a two dimensional integral. More precisely, we want to make a change of variables in an integral

$$\int_\Omega f(x)\,dx = \int_\Omega f(x_1, x_2)\,dx_1 dx_2, \tag{64.21}$$

where Ω is a given domain in $\mathbb{R}^2$ and the integration variable x runs over Ω. We assume that $g : \tilde{\Omega} \to \Omega$ is a one-to-one mapping of $y \in \tilde{\Omega}$ onto $x = g(y)$ in Ω that represents the change of variables. We shall prove that (64.21) with respect to x can be rewritten as an integral with respect to y in the form

$$\int_\Omega f(x)\,dx = \int_{\tilde{\Omega}} f(g(y))\,G(y)\,dy, \tag{64.22}$$

where $G(y)$ is defined as

$$G(y) = |\det g'(y)|.$$

That is $G(y)$ is the absolute value of the determinant of the Jacobian $g'(y)$ of $g(y)$. Formally, this gives $dx = |\det g'(y)|\,dy$ or $|\det g'(y)| = |\det \frac{dx}{dy}|$, and $|\det g'(y)|$ is the local change of area measure as we go from y-coordinates to x-coordinates. The change of variable formula can therefore be written

$$\int_\Omega f(x)\,dx = \int_{\tilde{\Omega}} f(g(y))|\det g'(y)|\,dy, \tag{64.23}$$

To prove this let $\tilde{\Omega}_i$ be a small subdomain of $\tilde{\Omega}$ and let $\Omega_i = g(\tilde{\Omega}_i)$ be the image of $\tilde{\Omega}_i$ under the mapping $x = g(y)$. If $g'(y)$ were constant over $\tilde{\Omega}_i$, and so $g(y)$ were linear on $\tilde{\Omega}_i$, then

$$d\Omega_i = |\det g'(y_i)| d\tilde{\Omega}_i,$$

where y_i is a point in $\tilde{\Omega}_i$, $d\Omega_i$ is the area of Ω_i, and $d\tilde{\Omega}_i$ is the area of $\tilde{\Omega}_i$. If $\{\tilde{\Omega}_i\}_{i=1}^n$ is a subdivision of $\tilde{\Omega}$ into subdomains $\tilde{\Omega}_i$ of maximal diameter d_n, we have

$$\begin{aligned}\int_\Omega f(x)\,dx &\approx \sum_i f(x_i) d\Omega_i \\ &\approx \sum_i f(g(y_i)|\det g'(y_i)| d\tilde{\Omega}_i \approx \int_{\tilde{\Omega}} f(g(y))|\det g'(y)|\,dy,\end{aligned}$$

where $x_i = g(y_i)$ and the approximations are bounded by d_n times Lipschitz constants of the functions $f(x)$, $f(g(y))$ and $|\det g'(y)|$. The change of variables formula (64.23) follows by passing to the limit as n tends to infinity and d_n tends to 0.

We summarize:

Theorem 64.5 (Change of variables) *Assume $y \to x = g(y)$ maps a domain $\tilde{\Omega}$ in $\mathbb{R}^2$ onto a domain Ω in $\mathbb{R}^2$, where the Jacobian of g is Lipschitz continuous and let $f : \Omega \to \mathbb{R}$ be Lipschitz continuous. Then*

$$\int_\Omega f(x)\,dx = \int_{\tilde{\Omega}} f(g(y))|\det g'(y)|\,dy, \tag{64.24}$$

EXAMPLE 64.6. Consider the mapping $x = g(y) = (2y_1 + y_2, y_1 - 2y_2)$ mapping the unit square $\tilde{\Omega} = [0,1] \times [0,1]$ onto the parallelogram Ω spanned by the vectors $(2,1)$ and $(1,-2)$. We have $\det g'(y) = -5$, and thus

$$\begin{aligned}\int_\Omega f(x)\,dx = \int_{\tilde{\Omega}} f(2y_1 + y_2, y_1 - 2y_2)\,|-5|\,dy \\ = 5\int_0^1\int_0^1 f(2y_1 + y_2, y_1 - 2y_2)\,dy.\end{aligned}$$

If $f(x) = x_2$ then

$$\int_\Omega f(x)\,dx = 5\int_0^1\int_0^1 (y_1 - 2y_2)\,dy = 5(\frac{1}{2} - 1) = -\frac{5}{2}.$$

64.10.1 Polar coordinates

A particularly important change of variables is from rectangular coordinates to polar coordinates,

$$(x_1, x_2) = (r\cos(\theta), r\sin(\theta))$$

where $x = (x_1, x_2) \in \mathbb{R}^2$ and $r \geq 0$, $0 \leq \theta < 2\pi$, see Fig. 64.8.

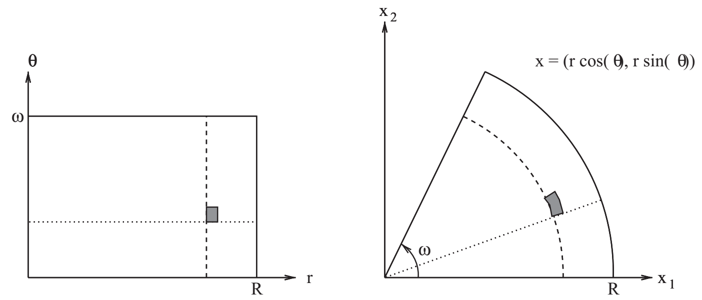


FIGURE 64.8. Polar coordinates

The Jacobian of the mapping $(r, \theta) \to (x_1, x_2)$ is given by

$$\frac{d(x_1, x_2)}{d(r, \theta)} = \begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix},$$

and

$$\det \frac{d(x_1, x_2)}{d(r, \theta)} = r(\cos^2(\theta) + \sin^2(\theta)) = r.$$

EXAMPLE 64.7. If $\Omega = \{x \in \mathbb{R}^2 : |x| \leq 1, x_1 \geq 0, x_2 \geq 0\}$ is the part of the unit circle in the positive quadrant, then the corresponding domain in polar coordinates takes the from $\tilde{\Omega} = \{(r, \theta) : 0 \leq r \leq 1, 0 \leq \theta \leq \frac{\pi}{2}\}$, and

$$\int_\Omega f(x_1, x_2)\, dx_1 dx_2 = \int_{\tilde{\Omega}} f(r\cos(\theta, r\sin(\theta))\, r dr\, d\theta.$$

In particular with $f(x) = 1$, we have

$$\begin{aligned} A(\Omega) &= \int_\Omega dx_1 dx_2 = \int_{\tilde{\Omega}} r dr\, d\theta \\ &= \int_0^{\frac{\pi}{2}} \int_0^1 r\, dr\, d\theta = \int_0^{\frac{\pi}{2}} \frac{1}{2}\, d\theta = \frac{\pi}{4}. \end{aligned}$$

We have now computed the area of a quarter of a unit disc to be equal to $\frac{\pi}{4}$, so the area of a unit disc is π. A basic result of mathematics!

EXAMPLE 64.8. Using polar coordinates, we have

$$\int_{\mathbb{R}^2} e^{-x_1^2 - x_2^2}\, dx = \int_0^{2\pi} \int_0^\infty e^{-r^2} r\, dr\, d\theta = 2\pi\big[-\frac{1}{2} e^{-r^2}\big]_0^\infty = \pi.$$

Since

$$\int_{\mathbb{R}^2} e^{-x_1^2 - x_2^2}\, dx = \int_{-\infty}^\infty e^{-x_1^2}\, dx_1 \int_{-\infty}^\infty e^{-x_2^2}\, dx_2,$$

we conclude that

$$\int_{-\infty}^\infty e^{-x^2}\, dx = \sqrt{\pi}. \qquad (64.25)$$

Evidently, we did something magical: although we do not know a primitive function to e^{-x^2} we are able to obtain an analytic expression for $\int_{-\infty}^\infty e^{-x^2}\, dx$.

Chapter 64 Problems

64.1. Compute with $\Omega = [0, 1] \times [0, 1]$ the unit square the integrals (a) $\int_\Omega (x_1 + x_2)\, dx$ (b) $\int_\Omega x_1 x_2\, dx$ (c) $\int_\Omega \frac{dx}{x_1 + x_2}$ (d) $\int_\Omega \exp(-x_1 x_2)\, dx$

64.2. Compute with $\Omega = \{(x_1, x_2) : 0 \leq x_1 \leq x_2 \leq 1\}$ the integrals (a) $\int_\Omega \frac{x_1}{x_2}\, dx$ (b) $\int_\Omega \exp^{2x_2}\, dx$ (c) $\int_\Omega \exp^{x_2^2}\, dx$

64.3. Change the order of integration in the following integrals

1. $\int_{1/2}^{1}\int_{0}^{1-x_1} f(x_1,x_2)\,dx_2dx_1$
2. $\int_0^1\int_0^{\sqrt{1-x_1^2}} f(x_1,x_2)\,dx_2dx_1$
3. $\int_0^1\int_{x_2-1}^{0} f(x_1,x_2)\,dx_1dx_2$
4. $\int_0^1\int_{1-x_1}^{1+x_1} f(x_1,x_2)\,dx_2dx_1$

64.4. Evaluate the following integrals:

1. $\int_\Omega (x_1^2+2x_2^3)\,dx$, with Ω a triangle with vertices $(0,0)$, $(1,0)$, $(0,1)$.
2. $\int_\Omega x_1^2x_2\,dx$, with $\Omega = \{x\in\mathbb{R}^2 : x_1^2+x_2^2\le 1,\ 0\le x_2$.
3. $\int_\Omega (x_1+x_2)\,dx$, with Ω the tetrahedron with vertices $(0,0)$, $(1,0)$, $(2,1)$, $(2,2)$.
4. $\int_\Omega |1-x_1-x_2|\,dx$, with Ω the unit square.

64.5. Find the volume under the graph of the following functions

1. $f(x)=e^{x_1}\cos(x_2)$, $0\le x_1\le 1$, $0\le x_2\le \frac{\pi}{2}$.
2. $f(x)=x_1^2e^{-x_1-x_2}$, $0\le x_1\le 1$, $0\le x_2\le 2$.
3. $f(x)=x_1^2x_2$, $0\le x_1\le 1$, $x_1+1\le x_2\le x_1+2$.
4. $f(x)=\sqrt{x_1^2-x_2^2}$, $x_1^2-x_2^2\ge 0$, $0\le x_1\le 1$.

64.6. A cylindrical hole of radius b is drilled symmetrically through a metal sphere of radius $a>b$. Find the volume of metal removed.

64.7. Evaluate
$$\int_\Omega (1-\frac{x_1^2}{a_1^2}-\frac{x_2^2}{a_2^2})^{3/2}\,dx$$
where Ω is the ellipse $\{x\in\mathbb{R}^2 : \frac{x_1^2}{a_1^2}+\frac{x_2^2}{a_2^2}\le 1\}$.

64.8. Evaluate
$$\int_\Omega \frac{x_1+x_2}{x_1^2}e^{x_1+x_2}\,dx,$$
where $\Omega=\{x\in\mathbb{R}^2 : x_2\le x_1\le 2-x_2,\ 0\le x_2\le 1\}$. Hint: Use the substitution $y_1=x_1+x_2$, $y_2=\frac{x_2}{x_1}$.

64.9. Compute the area of one petal of the rose $0\le r\le 3\sin(\theta)$ (polar coordinates).

64.10. Compute the area within the cardoid $r=1+\cos(\theta)$.

64.11. Compute the following double integrals:

1. $\int_\Omega x_1\exp(x_1x_2)\ dx$, for $\Omega=\{x : 0\le x_1\le 1, 0\le x_2\le 2\}$,
2. $\int_\Omega x_1x_2\exp(x_1+x_2)\ dx$, for $\Omega=\{x : 1\le x_1\le 2\le x_2\le 3\}$,
3. $\int_\Omega x\ dx$, for $\Omega=\{x : 0\le x_1\le 1, 0\le x_2\le 1\}$.

64.12. Compute the following double integrals:

1. $\int_\Omega \exp(-x_1)\, dx$, for $\Omega = \{x : 0 \le x_1 \le 1, |x_2| \le x_1\}$,
2. $\int_\Omega x_1 x_2 \|x\|\, dx$, for $\Omega = \{x : 0 \le x_1 \le 1, 1 \le x_2 \le 2\}$,
3. $\int_\Omega \frac{x_1}{1+x_2}\, dx$, for $\Omega = \{x : 0 \le x_1 \le 1, 0 \le x_2 \le 1 - x_1\}$.

64.13. Compute the following double integrals by changing variables:

1. $\int_\Omega \|x\|^2\, dx$, for $\Omega = \{x : x_1^2 + x_2^2 - 2x_1 - 2x_2 \le 0\}$,
2. $\int_\Omega x_1 x_2\, dx$, for $\Omega = \{x : 3x_1^2 + x_2^2 - 2x_1 \le 0\}$,
3. $\int_\Omega \exp(-\|x\|^2)\, dx$, for $\Omega = \mathbb{R}^2$.

65
Surface integrals

King Karl XII of Sweden (1682-1717) had an extraordinary talent for mathematics. He was by Swedenborg (the great Swedish Universal Genius, 1688-1772) considered equal if not better than Leibniz himself. King Karl XII could easily multiply large numbers without pen and paper, and proposed 64 as the right choice of basis of the natural numbers. Over night he constructed symbols and gave names to all the digits $0, 1, ..., 62, 63$. (from *The History of Sweden*, by Herman Lindquist).

65.1 Introduction

Previously, in Chapter *Curve integrals* we defined the notion of an integral computed over a curve or a curve integral. In this chapter, we use the same ideas to define an integral over a surface or a *surface integral.* We start with the surface integral representing *surface area.*

65.2 Surface area

Let S be a surface in $\mathbb{R}^3$ parameterized by the mapping $s : \Omega \to \mathbb{R}^3$, where Ω is a domain in $\mathbb{R}^2$ with coordinates $y = (y_1, y_2) \in \mathbb{R}^2$, so that $s = s(y) = (s_1(y), s_2(y), s_3(y))$. We define the *area* $A(S)$ of the surface S

as the following integral over the parameter domain Ω,

$$A(S) = \int_\Omega \|s'_{,1} \times s'_{,2}\| \, dy, \tag{65.1}$$

where

$$s'_{,1} = \begin{pmatrix} \frac{\partial s_1}{\partial y_1} \\ \frac{\partial s_2}{\partial y_1} \\ \frac{\partial s_3}{\partial y_1} \end{pmatrix}, \qquad s'_{,2} = \begin{pmatrix} \frac{\partial s_1}{\partial y_2} \\ \frac{\partial s_2}{\partial y_2} \\ \frac{\partial s_3}{\partial y_2} \end{pmatrix},$$

are the columns of the Jacobian

$$s' = \begin{pmatrix} \frac{\partial s_1}{\partial y_1} & \frac{\partial s_1}{\partial y_2} \\ \frac{\partial s_2}{\partial y_1} & \frac{\partial s_2}{\partial y_2} \\ \frac{\partial s_3}{\partial y_1} & \frac{\partial s_3}{\partial y_2} \end{pmatrix}.$$

Note all the coefficients are functions of $y \in \Omega$.

To motivate this definition, recall that the linearization of the mapping $s : \Omega \to \mathbb{R}^3$ at $\bar{y}$ is given by

$$y \to \hat{s}(y) = s(\bar{y}) + (y_1 - \bar{y}_1)s'_{,1}(\bar{y}) + (y_2 - \bar{y}_2)s'_{,2}(\bar{y}).$$

Consider a small square $R(\bar{y}, h) = [\bar{y}_1, \bar{y}_1 + h] \times [\bar{y}_2, \bar{y}_2 + h]$ in Ω of side length h and area h^2 with lower left-hand corner at the point $\bar{y} \in \Omega$. Here, we think of h as small. The linearization $\hat{s}(y)$ maps the square $R(\bar{y}, h)$ into a small parallelogram $P(s(\bar{y}), h)$ in the tangent plane of s through $s(\bar{y})$ spanned by the two vectors $s'_{,1}(\bar{y})$ and $s'_{,2}(\bar{y})$, with one of the corners of the parallelogram at $s(\bar{y})$. Recall (Chapter *Analytic geometry in* $\mathbb{R}^2$) that the area of the parallelogram spanned by two vectors a and b in $\mathbb{R}^2$ is equal to $\|a \times b\|$). So, the area of $P(s(\bar{y}), h)$ is equal to

$$\|s'_{,1}(\bar{y}) \times s'_{,2}(\bar{y})\|h^2.$$

The change of scale of area is thus $\|s'_{,1}(\bar{y}) \times s'_{,1}(\bar{y})\|$. A small piece (square) of area h^2 at $\bar{y} \in \Omega$ thus corresponds to a small piece of the surface S at $s(\bar{y})$ of area approximately $\|s'_{,1}(\bar{y}) \times s'_{,2}(\bar{y})\|h^2$, where the approximation improves as h gets smaller. Summing over all little pieces and letting h tend to zero, we are led to define the area $A(S)$ of the surface S by (65.1), which we write as

$$A(S) = \int_\Omega \|s'_{,1}(y) \times s'_{,2}(y)\| \, dy = \int_\Omega \|s'_{,1} \times s'_{,2}\| \, dy = \int_S ds.$$

We thus write $ds = \|s'_{,1} \times s'_{,2}\| \, dy$, which expresses the change of scale. Of course, we assume that $\|s'_{,1} \times s'_{,2}\|$ is Lipschitz continuous to guarantee that the integral exists.

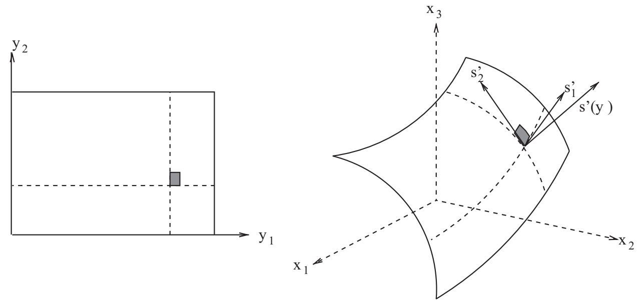


FIGURE 65.1. The surface area scale

EXAMPLE 65.1. Consider the surface S of a sphere of radius one centered at the origin. We describe this using spherical coordinates,

$$x = s(y_1, y_2) = (\sin(y_2)\cos(y_1), \sin(y_2)\sin(y_1), \cos(y_2))^\top,$$

where $0 \le y_1 < 2\pi$, $0 \le y_2 < \pi$, see Fig. 66.3. We have

$$\begin{aligned} s'_{,1} &= (-\sin(y_2)\sin(y_1), \sin(y_2)\cos(y_1), 0)^\top, \\ s'_{,2} &= (\cos(y_2)\cos(y_1), \cos(y_2)\sin(y_1), -\sin(y_2))^\top, \end{aligned} \tag{65.2}$$

and thus by a direct computation $\|s'\| = \sin(y_2)$. We compute

$$A(S) = \int_0^{2\pi}\int_0^{\pi} \sin(y_2)\, dy_2\, dy_1 = \int_0^{2\pi} 2\, dy_1 = 4\pi,$$

and thus conclude that the surface area of a sphere of radius 1 is equal to 4π.

EXAMPLE 65.2. We compute the area $A(S)$ of the surface S given by $s(y_1, y_2) = (2y_1y_2, y_1^2, 2y_2^2)$ with $0 \le y_1, y_2 \le 1$. We have

$$s'(y) = (2y_2, 2y_1, 0) \times (2y_1, 0, 4y_2) = 4(2y_1y_2, -2y_2^2, -y_1^2)$$

so that $\|s'(y)\| = 4(y_1^2 + 2y_2^2)$, and thus

$$A(S) = \int_0^1\int_0^1 4(y_1^2 + 2y_2^2)\, dy_1\, dy_2 = 4(\frac{1}{3} + \frac{2}{3}) = 4.$$

65.3 The surface area of a function of two variables

In the case S is given as the graph of a function $f : \Omega \to \mathbb{R}$, so that $s(y_1, y_2) = (y_1, y_2, f(y_1, y_2))$, then

$$A(S) = \int_S ds = \int_\Omega \|s'_{,1} \times s'_{,2}\|\, dy = \int_\Omega \sqrt{1 + f_{,1}^2 + f_{,2}^2}\, dy_1 dy_2, \tag{65.3}$$

where $f_{,i}$ denotes the partial derivative of f with respect to y_i. This follows from

$$s'_{,1} \times s'_{,2} = (1, 0, f_{,1}) \times (0, 1, f_{,2}) = (-f_{,1}, -f_{,2,}, 1).$$

EXAMPLE 65.3. The surface S of a hemisphere of radius 1 and centered at the origin is given by $s(y_1, y_2) = (y_1, y_2, \sqrt{1 - y_1^2 - y_2^2})$ with $y \in \Omega = \{y \in \mathbb{R}^2 : y_1^2 + y_2^2 \leq 1\}$. We have

$$\begin{aligned} A(S) &= \int_\Omega \sqrt{1 + f_{,1}^2 + f_{,2}^2}\, dy_1 dy_2 = \int_\Omega \frac{1}{\sqrt{1 - y_1^2 - y_2^2}}\, dy \\ &= \int_0^{2\pi} \int_0^1 \frac{1}{\sqrt{1 - r^2}}\, r\, dr\, d\theta = 2\pi \big[-\sqrt{1 - r^2}\big]_0^1 = 2\pi. \end{aligned} \tag{65.4}$$

We retrieve the above result that the surface area of a sphere of radius 1 is equal to 4π.

65.4 Surfaces of revolution

Surfaces of revolution occur in many practical applications. To generate a surface of revolution, we let $f : [a, b] \to \mathbb{R}$ be a given positive function and consider the surface S represented by $s(x_1, x_2) = (x_1, f(x_1)\cos(x_2), f(x_1)\sin(x_2))$ with $a \leq x_1 \leq b$ and $0 \leq x_2 < 2\pi$, see Fig. 65.2. We use (x_1, x_2) as reference coordinates instead of (y_1, y_2). We have

$$s'_{,1} \times s'_{,2} = (1, f'(x_1)\cos(x_2), f'(x_1)\sin(x_2)) \times (0, -f(x_1)\sin(\theta), f(x_1)\cos(\theta))$$

and thus by a direct computation

$$\|s'_{,1} \times s'_{,2}\| = f(x_1)\sqrt{1 + (f'(x_1))^2}. \tag{65.5}$$

The area $A(S)$ of S is given by:

$$\begin{aligned} A(S) &= \int_0^{2\pi} \int_a^b f(x_1)\sqrt{1 + (f'(x_1))^2}\, dx_1 d\theta \\ &= 2\pi \int_a^b f(x_1)\sqrt{1 + (f'(x_1))^2}\, dx_1. \end{aligned} \tag{65.6}$$

EXAMPLE 65.4. Consider the surface S of a parabolic reflector obtained by rotating the curve $f(x_1) = \sqrt{x_1}$ around the x_1-axis between $x_1 = 0$ and $x_1 = 1$. We have

$$A(S) = 2\pi \int_0^1 \sqrt{x_1}\sqrt{1 + \frac{1}{4x_1}}\, dx_1 = \pi \int_0^1 \sqrt{4x_1 + 1}\, dx_1 = \frac{\pi}{6}(5^{3/2} - 1).$$

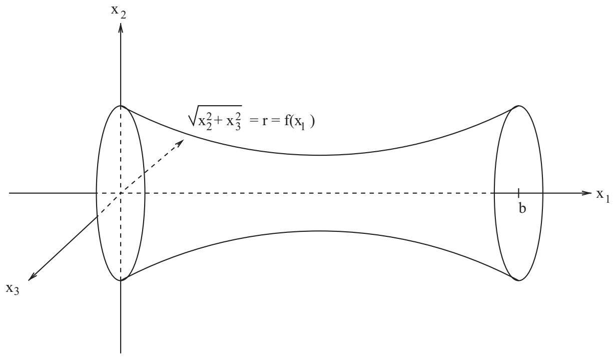


FIGURE 65.2. A surface of revolution

65.5 Independence of parameterization

We shall prove that if $t : \tilde{\Omega} \to \Omega$ is a one-to-one mapping of $\eta \in \tilde{\Omega} \subset \mathbb{R}^2$ onto $y = t(\eta) \in \Omega$, and $r(\eta) = s(t(\eta))$ maps $\tilde{\Omega}$ onto S, then

$$\int_S ds = \int_{\tilde{\Omega}} \|r'_{,1} \times r'_{,2}\| \, d\eta = \int_{\Omega} \|s'_{,1} \times s'_{,2}\| \, dy. \tag{65.7}$$

This shows that the surface area of the surface S is independent of the parametrization of S.

We need to show that with $y = t(\eta)$, we have

$$\|r'_{,1}(\eta) \times r'_{,2}(\eta)\| = \|s'_{,1}(y) \times s'_{,2}(y)\| \, |\det t'(\eta)|, \tag{65.8}$$

where $|\det t'|$ is the determinant of the Jacobian $t'(\eta)$ of $t(\eta)$. This follows after a lengthy computation that starts with differentiating $r(\eta) = s(t(\eta))$ using the Chain rule.

65.6 Surface integrals

Let $S = s(\Omega)$ be a surface in $\mathbb{R}^3$ parameterized by the mapping $s : \Omega \to \mathbb{R}^3$, where Ω is a domain in $\mathbb{R}^2$, and let $u : S \to \mathbb{R}$ be a real-valued function defined on S. We assume that u, s and $\|s'_{,1} \times s'_{,2}\|$ are Lipschitz continuous. We define the integral of u over S to be

$$\int_S u \, ds = \int_{\Omega} u(s(y)) \|s'_{,1}(y) \times s'_{,2}(y)\| \, dy. \tag{65.9}$$

Example 65.5. Let $S = s(\Omega)$ be the "dome" given by $s(y_1, y_2) = (y_1, y_2, 1 - y_1^2 - y_2^2)$ and $\Omega = \{y \in \mathbb{R}^2 : y_1^2 + y_2^2 \leq 1\}$, and $u(x) =$

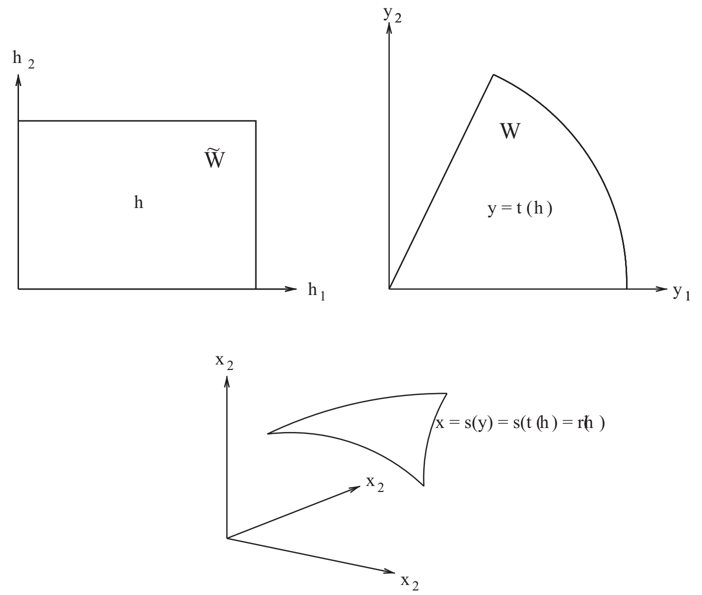


FIGURE 65.3. Reparametrization $r(\eta) = s(t(\eta))$ of a surface given by $s(y)$.

$(5x_1^2 + 5x_2^2 + x_3)^{1/2}$, so that $u(s(y)) = (5y_1^2 + 5y_2^2 + 1 - y_1^2 - y_2^2)^{1/2} = (1 + 4y_1^2 + 4y_2^2)^{1/2}$. We compute

$$\|s'_{,1}(y) \times s'_{,2}(y)\| = \|(1, 0, -2y_1) \times (0, 1, -2y_2)\| = (1 + 4y_1^2 + 4y_2^2)^{1/2},$$

and get using polar coordinates:

$$\int_S u\,ds = \int_\Omega u(s(y))\|s'_{,1}(y) \times s'_{,2}(y)\|\,dy$$
$$= \int_\Omega (1 + 4y_1^2 + 4y_2^2)^{1/2}(1 + 4y_1^2 + 4y_2^2)^{1/2}\,dy = 2\pi \int_0^1 (1 + 4r^2)r\,dr = \frac{3}{2}.$$

65.7 Moment of inertia of a thin spherical shell

The *moment of inertia* of a thin sphere $S = \{\|x\| = 1\}$ of (uniformly distributed) total mass m about the x_3-axis, is equal to

$$I = \frac{m}{4\pi} \int_S (x_1^2 + x_2^2)\,ds. \tag{65.10}$$

If the sphere rotates with angular speed ω around the x_3 axis, then the total *kinetic energy* is equal to

$$E = \frac{1}{2}\frac{m}{4\pi} \int_S \omega^2(x_1^2 + x_2^2)\,ds = \frac{1}{2}\omega^2 I. \tag{65.11}$$

Using spherical coordinates to compute gives

$$I = \frac{2m}{3}. \tag{65.12}$$

Chapter 65 Problems

65.1. (a) Verify that $\|s'_{,1}(y) \times s'_{,2}(y)\| = \sin(y_2)$ in (65.2).
(b) Verify (65.5).
(c) Prove (65.8).

65.2. Find out which famous building is defined by the following *MATLAB*© code, and compute the surface area of its roof.

```
r=0:.1:1;
v=0:pi/20:2*pi;
[R,V]=meshgrid(r,v);
surf(10*cos(V),10*sin(V),R.*(5+cos(V).^2-sin(V).^2))
hold on
surf(10*R.*cos(V),10*R.*sin(V),5+(R.*cos(V)).^2-(R.*sin(V)).^2)
hold off
axis('equal')
```

65.3. Another famous building. What does it take to repaint it?

```
w=0:pi/20:3*pi/4;
v=0:pi/20:2*pi;
[W,V]=meshgrid(w,v);
h=surf(sin(W).*cos(V),sin(W).*sin(V),cos(W));
set(h,'FaceColor',[1 1 1])
axis('equal')
```

65.4. Motivate (65.11), and prove (65.12).

65.5. (a) Consider the surface $S = \{x : x = y_1 a + y_2 b + (1 - y_1 - y_2)c, y \in T\}$, where $a, b, c \in \mathbb{R}^3$ and $T = \{y \in \mathbb{R}^2 : y_1 + y_2 \leq 1, y_i \geq 0, i = 1, 2\}$. Give a geometric description of S and compute its area.
(b) Find a parametrization of the form $x = My + b$ of the (flat) triangular surface S with corners in $(1, 0, 0)$, $(0, 0, 3)$ and $(0, 3, -9)$, with parameter domain T as in (a), where b is a 3-vector and M a 3-by-2 matrix.
(b) Compute the area of S. Does the area depend on b? Interpret!
(c) Compute $\int_S (x_1 + 2x_2)\, dS$

65.6. Compute (a) $\int_S dS$ (b) $\int_S f(x)\, dS$ where $S = \{x : x = My, y \in Q\}$, Q is the unit square in $\mathbb{R}^2$ and M is the 3-by-2 matrix with columns $(1, 0, 1)^\top$ and

$(0,1,2)^\top$, and $f(x) = x_3$. Also, plot the surface S and interpret (a) as the area of S. Compare the computation of (a) with the method for computing the area of a parallelogram using the cross product in Linear algebra.

65.7. Compute (a) $\int_S dS$ (b) $\int_S x_2\, dS$ where $S = \{x : x = y_1(1-y_2)(1,0,0) + (1-y_1)(1-y_2)(1,2,0) + (1-y_1)y_2(0,1,1) + y_1y_2(0,0,3), 0 \le y_i \le 1, i = 1,2\}$. Plot the surface and describe its geometry.

65.8. Compute (a) $\int_S dS$ (b) $\int_S x_1x_2\, dS$ where $S = \{(y_1, y_2, y_1y_2) : 0 \le y_i \le 1, i = 1,2\}$.

65.9. Consider for given $r > 0$ and $h > 0$ the surface $S = \{x : x = (r\cos(v), r\sin(v), z),\ 0 \le v \le 2\pi,\ 0 \le z \le h\}$. (a) Give a geometrical description of S, and give corresponding parameterizations of the surfaces (b) $S = \{x \in \mathbb{R}^3 : x_2^2 + x_3^2 = 4,\ |x_1| \le 5\}$ (c) $S = \{x \in \mathbb{R}^3 : x_2^2 + 4x_3^2 = 4,\ 0 \le x_1 \le x_2^2 + x_3^2\}$.

65.10. Compute $\int_S (x_1+x_2+x_3)\, dS$ where $S = \{(x_1,x_2,x_3) : x_1 = y_1\cos(y_2), x_2 = y_1\sin(y_2), x_3 = y_1(\cos(y_2)+\sin(y_2))\}$.

65.11. Compute $\int_S (x_1,x_2,x_3)\cdot n\, ds$ if S is the boundary of $\Omega = \{x : x_1+x_2+x_3 \le 1, x_i \ge 0, i = 1,2,3\}$.

65.12. Compute $\int_S \frac{(x_1,x_2,x_3)}{\|x\|^2} \cdot n\, dS$ for the cylindrical shell $S = \{x \in \mathbb{R}^3 : x_1^2 + x_2^2 = 1, -a \le x_3 \le a\}$, and the corresponding limit as $a \to \infty$.

65.13. Compute the moment of inertia of the cylindrical shell $S = \{x \in \mathbb{R}^3 : x_1^2 + x_2^2 = 1, -1 \le x_3 \le 1\}$ with respect to the x_1-axis.

65.14. Compute $\int_S (x_1, 0, x_3)\cdot n\, dS$ where $S = \{(y_1+y_2, y_1^2 - y_2^2, y_1y_2) : 0 \le y_1 \le 1, 0 \le y_2 \le 1\}$, and n is the normal to S with $n_3 < 0$.

65.15. Compute the area of the torus (donut) in $\mathbb{R}^3$ given by

$$s(y_1,y_2) = \big((a + b\cos(y_2))\cos(y_1), (a + b\cos(y_2))\sin(y_1), b\sin(y_2)\big)$$

with $a > b$ constants and $0 \le y_1, y_2 < 2\pi$.

65.16. Plot and compute the area of the surface $S = \{(r\cos(v), r\sin(v), v) : 1 \le r \le 2,\ 0 \le v \le 4\pi\}$. In what type of buildings can one find constructions like this?

65.17. Describe/plot the surfaces (of rotation, if you wish) (a) $x_1^2+x_2^2 = x_3^2$, $x_3 > 0$ (b) $5 + x_1^2 + x_2^2 = x_3^2 \le 9$, $x_3 > 0$ and compute its area.

66
Multiple integrals

We met weekly, (sometimes at Dr Goddard's lodgings, sometimes at the Mitre in Wood Street near-by) at a certain hour, under a certain penalty, and a weekly contribution for the charge of experiments, with certain rules agreed among us. There, to avoid being diverted to other discourses and for some other reasons, we barred all discussion of Divinity, of State Affairs, and of news (other than what concerned our business of philosophy) confining ourselves to philosophical inquiries, and related topics; as medicine, anatomy, geometry, astronomy, navigation, statics, mechanics, and natural experiments. (Wallis about the formation of the Royal Society)

66.1 Introduction

We now consider *triple integrals* over domains in $\mathbb{R}^3$ and more generally *multiple integrals* over domains in $\mathbb{R}^n$ with $n > 3$.

66.2 Triple integrals over the unit cube

A triple integral over the unit cube $Q = \{x \in \mathbb{R}^3 : 0 \leq x_i \leq 1,\, i = 1, 2, 3\}$ of a Lipschitz continuous function $f : Q \to \mathbb{R}$ takes the form

$$\int_Q f(x)\, dx = \int_0^1 \int_0^1 \int_0^1 f(x_1, x_2, x_3)\, dx_1\, dx_2\, dx_3.$$

This can be computed by iterated integration in any order, for example,

$$\int_Q f(x)\,dx = \int_0^1 \Big(\int_0^1 \big(\int_0^1 f(x_1, x_2, x_3)\,dx_3\Big)\,dx_2\Big)\,dx_1.$$

The definition of the integral and the verification of the iterated integration formula is a direct generalization of the corresponding steps in the case of a double integral over the unit square.

EXAMPLE 66.1. We compute the integral of $x_1^2 x_2 e^{x_1 x_2 x_3}$ over the unit cube Q,

$$\int_Q x_1^2 x_2 e^{x_1 x_2 x_3}\,dx = \int_0^1 \int_0^1 \Big(\int_0^1 x_1^2 x_2 e^{x_1 x_2 x_3}\,dx_3\Big) dx_1\,dx_2$$
$$= \int_0^1 \int_0^1 \Big[x_1 e^{x_1 x_2 x_3}\Big]_{x_3=0}^{x_3=1} dx_1 dx_2 = \int_0^1 \int_0^1 x_1 (e^{x_1 x_2} - 1)\,dx_1 dx_2.$$

which leaves a double integral that we know how to handle.

66.3 Triple integrals over more general domains in $\mathbb{R}^3$

Let $\Omega = \{x \in \mathbb{R}^3 : \gamma_2(x_1, x_2) \le x_3 \le \gamma_1(x_1, x_2),\ (x_1, x_2) \in \omega\}$, where ω is a domain in $\mathbb{R}^2$ and $\gamma_1 : \omega \to \mathbb{R}$ and $\gamma_2 : \omega \to \mathbb{R}$ are given functions of (x_1, x_2), see Fig. 66.1. Let $f : \Omega \to \mathbb{R}$ Lipschitz continuous. We define the triple integral of $f(x)$ over Ω by

$$\int_\Omega f(x)\,dx = \int_\omega \Big(\int_{\gamma_2(x_1,x_2)}^{\gamma_1(x_1,x_2)} f(x_1, x_2, x_3)\,dx_3\Big) dx_1 dx_2$$

via iterated integration first with respect to x_3 and then with respect to $(x_1, x_2) \in \omega$.

Expanding the double integral over ω into two one-dimensional integrals, assuming $\omega = \{(x_1, x_2, x_3) : \alpha_2 \le x_1 \le \alpha_1,\ \beta_2(x_1) \le x_2 \le \beta_1(x_1)\}$, we have

$$\int_\Omega f(x)\,dx = \int_{\alpha_2}^{\alpha_1} \Big(\int_{\beta_2(x_1)}^{\beta_1(x_1)} \Big(\int_{\gamma_2(x_1,x_2)}^{\gamma_1(x_1,x_2)} f(x)\,dx_3\Big)\,dx_2\Big)\,dx_1$$
$$= \int_{\alpha_2}^{\alpha_1} \Big(\int_{\omega(x_1)} f(x_1, x_2, x_3)\,dx_2 dx_3\Big)\,dx_1,$$

where $\omega(x_1) = \{(x_2, x_3) : \beta_2(x_1) \le x_2 \le \beta_1(x_1),\ \gamma_2(x_1, x_2) \le x_3 \le \gamma_1(x_1, x_2)\}$ is the cross-section of the domain Ω with a plane with fixed x_1-coordinate. This way of splitting a triple integral into an one-dimensional integral of double integrals over domain cross-sections corresponds to cutting a piece of bread or ham into slices.

We may define triple integrals similarly for more general domains by dividing the domain suitably into pieces.

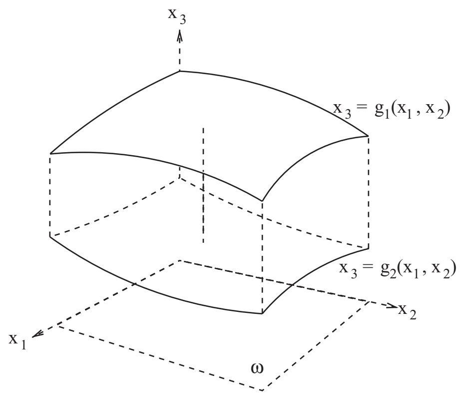


FIGURE 66.1. Integration over a volume by first integrating in the x_3-direction.

66.4 The volume of a three-dimensional domain

We define the volume $V(\Omega)$ of a domain Ω in $\mathbb{R}^3$ as

$$V(\Omega) = \int_\Omega dx,$$

i.e. by integrating $f(x) = 1$ over $x \in \Omega$. If $\Omega = \{x \in [0,1] \times [0,1] \times \mathbb{R} : \gamma_2(x_1, x_2) \leq x_3 \leq \gamma_1(x_1, x_2)\}$, then

$$\begin{aligned} V(\Omega) &= \int_0^1 \int_0^1 \Big(\int_{\gamma_2(x_1,x_2)}^{\gamma_1(x_1,x_2)} dx_3 \Big)\, dx_1 dx_2 \\ &= \int_0^1 \int_0^1 \Big(\gamma_1(x_1) - \gamma_2(x_1) \Big)\, dx_1\, dx_2, \end{aligned}$$

which conforms to the previous formula of the volume between the surfaces $\gamma_1(x_1, x_2)$ and $\gamma_2(x_1, x_2)$ as the integral of the difference $\gamma_1(x_1, x_2) - \gamma_2(x_1, x_2)$.

EXAMPLE 66.2. The volume of the pyramid Ω with corners at $(0,0,0)$, $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$ described as $\{x \in \mathbb{R}^3 : 0 \leq x_1 + x_2 + x_3 \leq$

$1,\ x_1x_2, x_3 \geq 0\}$, can be computed with $\omega = \{(x_1, x_2) : 0 \leq x_1 \leq 1, 0 \leq x_2 \leq 1 - x_1\}$ as follows

$$\begin{aligned} V(\Omega) &= \int_\omega \Big(\int_0^{1-x_1-x_2} dx_3\Big) dx_1\, dx_2 = \int_\Omega dx \\ &= \int_0^1 \Big(\int_0^{1-x_1} \Big(\int_0^{1-x_1-x_2} dx_3\Big)\, dx_2\Big) dx_1 \\ &= \int_0^1 \Big(\int_0^{1-x_1} (1 - x_1 - x_2)\, dx_2\Big) dx_1 = \int_0^1 (1 - x_1)^2/2\, dx_1 = \frac{1}{6}, \end{aligned}$$

which agrees with the earlier computation giving the volume of a pyramid as $\frac{1}{3}Bh$, where B is the area of the base and h the height, see Fig. 66.2.

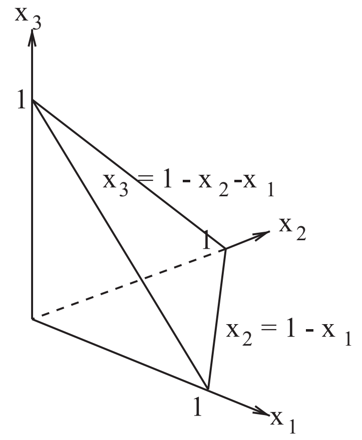


FIGURE 66.2. Integration over a pyramid.

66.5 Triple integrals as limits of Riemann sums

We may also define integrals over domains in $\mathbb{R}^3$ as limits of Riemann sums

$$\int_\Omega f(x)\, dx = \lim_{N\to\infty} \sum_{i=1}^{N} f(x_i) d\Omega_i, \tag{66.1}$$

where $\{\Omega_i\}_{i=1}^N$ is a subdivision of the given domain Ω into pieces Ω_i with volume $V(\Omega_i) \leq d_N$ and quadrature points $x_i \in \Omega_i$, where d_N tends to 0 as N tends to infinity. The error estimate (64.20) for double integrals generalizes directly to three dimensions.

66.6 Change of variables in a triple integral

We next prove an analog of the change of variable formula for two-dimensional integrals. We thus want to make a change of variables in an integral

$$\int_\Omega f(x)\,dx = \int_\Omega f(x_1, x_2, x_3)\,dx_1 dx_2 dx_3, \tag{66.2}$$

where Ω is a given domain in $\mathbb{R}^3$ and the integration variable x runs over Ω. If $g : \tilde{\Omega} \to \Omega$ is a one-to-one mapping of $y \in \tilde{\Omega}$ onto $x = g(y)$ in Ω, we have the following change of variables formula

$$\int_\Omega f(x)\,dx = \int_{\tilde{\Omega}} f(g(y))|\det g'(y)|\,dy, \tag{66.3}$$

where $|\det g'(y)|$ is the modulus of the determinant of the Jacobian $g'(y)$ of $g(y)$. Formally, we write $dx = |\det g'(y)|dy$ or $|\det g'(y)| = |\det \frac{dx}{dy}|$, and $|\det g'(y)|$ is the local change of volume measure as we go from y-coordinates to x-coordinates.

To prove this, let $\tilde{\Omega}_i$ be a small subdomain of $\tilde{\Omega}$ and let $\Omega_i = g(\tilde{\Omega}_i)$ be the image of $\tilde{\Omega}_i$ under the mapping $x = g(y)$. If $g'(y)$ were constant and$g(y)$ were linear over $\tilde{\Omega}_i$, then

$$d\Omega_i = |g'(y_i)|d\tilde{\Omega}_i, \tag{66.4}$$

where y_i is a point in $\tilde{\Omega}_i$, $d\Omega_i$ is the area of Ω_i, and $d\tilde{\Omega}_i$ is the area of $\tilde{\Omega}_i$. Thus,

$$\begin{aligned}\int_\Omega f(x)\,dx &\approx \sum_i f(x_i)d\Omega_i \\ &\approx \sum_i f(g(y_i)|\det g'(y_i)|d\tilde{\Omega}_i \approx \int_{\tilde{\Omega}} f(g(y))|\det g'(y)|\,dy,\end{aligned}$$

where $x_i = g(y_i)$ and $\{\tilde{\Omega}_i\}_{i=1}^N$ is a subdivision of $\tilde{\Omega}$ of maximal diameter d_N. Assuming now that $f(x)$, $f(g(y))$ and $|\det g'(y)|$ are Lipschitz continuous, the formula (66.3) follows by passing to the limit as d_N tends to 0.

66.6.1 Spherical coordinates

As a particular important change of variables, we consider spherical coordinates,

$$(x_1, x_2, x_3) = (r\sin(\varphi)\cos(\theta), r\sin(\varphi)\sin(\theta), r\cos(\varphi)),$$

where $x = (x_1, x_2, x_3) \in \mathbb{R}^3$ and $r \geq 0$, $0 \leq \theta < 2\pi$, $0 \leq \varphi < \pi$, see Fig. 66.3.

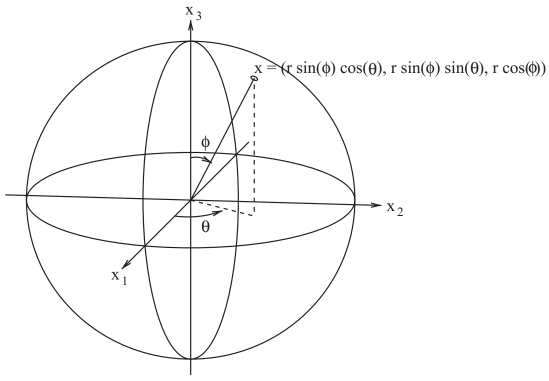


FIGURE 66.3. Spherical coordinates

The Jacobian of the mapping $(r,\theta,\varphi) \to (x_1,x_2,x_3)$ is equal to

$$\frac{d(x_1,x_2,x_3)}{d(r,\theta,\varphi)} = \begin{pmatrix} \sin(\varphi)\cos(\theta) & -r\sin(\varphi)\sin(\theta) & r\cos(\varphi)\cos(\theta) \\ \sin(\varphi)\sin(\theta) & r\sin(\varphi)\cos(\theta) & r\cos(\varphi)\sin(\theta) \\ \cos(\varphi) & 0 & -r\sin(\varphi) \end{pmatrix}$$

and by a direct computation

$$|\det \frac{d(x_1,x_2,x_3)}{d(r,\theta,\varphi)}| = r^2\sin(\varphi). \tag{66.5}$$

The change of variables formula from Cartesian x-coordinates to spherical coordinates takes the form

$$\int_\Omega f(x)\,dx = \int_{\tilde\Omega} f\big(r\sin(\varphi)\cos(\theta), r\sin(\varphi)\sin(\theta), r\cos(\varphi)\big)r^2\sin(\varphi)\,dr\,d\theta\,d\varphi,$$

where it is understood that $\tilde\Omega$ is a subdomain of $\{(r,\theta,\varphi) : 0 \le r,\ 0 \le \theta \le 2\pi,\ 0 \le \varphi \le \pi\}$, so that $(r,\theta,\varphi) \to x$ is a one-to-one mapping of $\tilde\Omega$ onto Ω.

EXAMPLE 66.3. The unit ball $B = \{x \in \mathbb{R}^3 : |x| \le 1,\}$ is described in spherical coordinates as $\tilde B = \{(r,\theta,\varphi) : 0 \le r \le 1,\ 0 \le \theta < 2\pi,\ 0 \le \varphi < \pi\}$. The volume $V(B)$ of B is given by

$$\begin{aligned} B = \int_B dx = \int_{\tilde B} dr\,d\theta\,d\varphi &= \int_0^\pi \int_0^{2\pi} \int_0^1 r^2\sin(\varphi)\,dr\,d\theta\,d\varphi \\ &= \int_0^\pi \sin(\varphi)\,d\varphi \int_0^{2\pi} d\theta \int_0^1 r^2\,dr = 2\cdot 2\pi\frac{1}{3} = \frac{4\pi}{3}. \end{aligned}$$

Note the way the triple integral splits into a product of three one-dimensional integrals because the limits of integration are fixed numbers in all the coordinate directions and the function to be integrated is a product of functions of the individual variables.

We have shown that the volume of the unit ball in $\mathbb{R}^3$ to be equal to $\frac{4\pi}{3}$. Another basic result of Calculus!

66.7 Solids of revolution

To generate a *solid of revolution*, we let $f : [a,b] \to \mathbb{R}$ be a given (positive) function and consider the body B in $\mathbb{R}^3$ represented by $s(x_1, r, \theta) = (x_1, r\cos(\theta), r\sin(\theta))$ with $a \le x_1 \le b$, $0 \le \theta < 2\pi$ and $0 \le r \le f(x_1)$, see Fig. 66.4. We have

$$\frac{d(x_1,x_2,x_3)}{d(x_1,r,\theta)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & \sin(\theta) \\ 0 & -r\sin(\theta) & r\cos(\theta) \end{pmatrix}$$

and thus by a direct computation

$$|\det \frac{d(x_1,x_2,x_3)}{d(x_1,r,\theta)}| = r.$$

The coordinate system (x_1, r, θ) is an example of so called cylindrical coordinates suitable for data with rotational symmetry.

The volume $V(B)$ of B is given by:

$$V(B) = \int_0^{2\pi} \int_a^b \int_0^{f(x_1)} r\, dr dx_1 d\theta = \pi \int_a^b f^2(x_1)\, dx_1. \qquad (66.6)$$

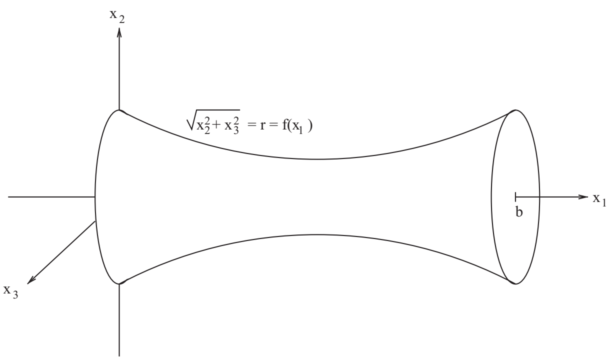


FIGURE 66.4. A solid of revolution

EXAMPLE 66.4. Consider the body B obtained by rotating the parabola $f(x_1) = \sqrt{x_1}$ around the x_1-axis between $x_1 = 0$ and $x_1 = 1$. We have

$$V(B) = \pi \int_0^1 x_1\, dx_1 = \frac{\pi}{2}.$$

EXAMPLE 66.5. The center of mass $\bar{x}$ of a body B of revolution obtained rotating a curve $f(x_1)$ around the x_1-axis from $x_1 = a$ to $x_1 = b$ is given by $\bar{x}_2 = \bar{x}_3 = 0$ (rotational symmetry) and

$$\bar{x}_1 = \frac{\pi}{V(B)} \int_a^b x_1 f^2(x_1)\, dx_1.$$

66.8 Moment of inertia of a ball

The *moment of inertia* about the x_3-axis of the ball $B = \{\|x\| = 1\}$ of (uniformly distributed) total mass m, is equal to

$$I = \frac{m}{V(B)} \int_B (x_1^2 + x_2^2)\, dx. \tag{66.7}$$

If the ball rotates with angular speed ω around the x_3 axis, then the total *kinetic energy* is equal to

$$E = \frac{1}{2}\frac{m}{V(B)} \int_B \omega^2 (x_1^2 + x_2^2)\, dx = \frac{1}{2}\omega^2 I. \tag{66.8}$$

Using spherical coordinates gives

$$I = \frac{2m}{5}. \tag{66.9}$$

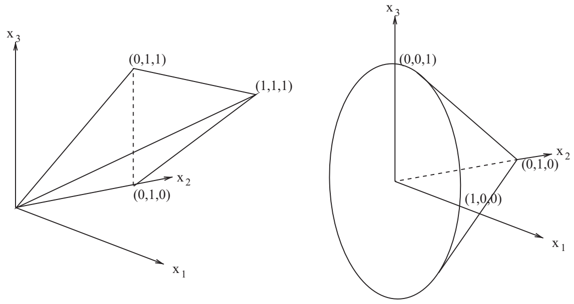
Chapter 66 Problems

66.1. Motivate (66.8) and prove (66.9).

66.2. Verify (66.5).

66.3. Compute the following triple integrals:

1. $\int_\Omega \|x\|^2\, dx$, for $\Omega = \{x \in \mathbb{R}^3 : 0 \le x_i \le 1, i = 1, 2, 3\}$,
2. $\int_\Omega \exp(x_1 + x_2 + x_3)\, dx$, for $\Omega = \{x \in \mathbb{R}^3 : 0 \le x_i \le 1, i = 1, 2, x_3 \le x_1 + x_2\}$,
3. $\int_\Omega 1/\|x\|^2\, dx$, for $\Omega = \{x \in \mathbb{R}^3 : 1 \le \|x\| \le 2\}$.



66.4. Compute with domains Ω as in Fig. 66.4 $\int_\Omega (1-x_2)\,dx$.

66.5. Compute for $\Omega = \{(x_1,x_2,x_3) : x_1^2 + x_2^2 \le 1, |x_3| \le 1\}$

1. $\int_\Omega \frac{dx}{\|x\|^2}$
2. The moment of inertia of Ω with respect to the x_3-axis.
3. The moment of inertia of Ω with respect to the x_2-axis.

66.6. Compute the following multiple integrals:

1. $\int_\Omega \frac{\exp(-\|x\|)}{\|x\|}\,dx$, for $\Omega = \{x \in \mathbb{R}^3 : \|x\| > 1\}$,
2. $\int_\Omega x_1 + x_2 + x_3 + x_4\ dx$, for $\Omega = \{x \in \mathbb{R}^4 : 0 \le x_i \le 1, i = 1,2,3,4\}$,
3. $\int_\Omega x_1 + \ldots + x_n\ dx$, for $\Omega = \{x \in \mathbb{R}^n : 0 \le x_i \le 1, i = 1,\ldots,n\}$.

66.7. Compute the following multiple integrals:

1. $\int_\Omega x\ dx$,
2. $\int_\Omega \|x\|\ dx$,
3. $\int_\Omega \|x\|^2\ dx$,

where $\Omega = \{x \in \mathbb{R}^3 : \|x\| \le 1\}$.

66.8. Try to generalize the result in the previous exercise to $\mathbb{R}^n$, denoting the area of the unit sphere, $\{x \in \mathbb{R}^n : \|x\| = 1\}$, by S_n.

66.9. Compute the integral $\int_{\mathbb{R}^2} \exp(-\|x\|^2)\ dx$ and use the result to compute $\int_{\mathbb{R}^n} \exp(-\|x\|^2)\ dx$.

66.10. Find the moment of inertia of a unit cube with respect to its diagonal.

66.11. Let E_y be the domain in $\mathbb{R}^n$ where the absolute value of $f : \mathbb{R}^n \to \mathbb{R}$ is larger than y, i.e. $E_y = \{x \in \mathbb{R}^n : |f(x)| > y\}$, and let $g(y)$ be the volume (size, measure) of this domain, i.e. $g(y) = \int_{E_y} dx$. Show, by changing the order of integration, that

$$\int_{\mathbb{R}^n} |f(x)|\ dx = \int_0^\infty g(y)\ dy.$$

67
Gauss' theorem and Green's formula in $\mathbb{R}^2$

Mathematics at its best: it looks impressive (incomprehensible), but is trivial for anyone who understands the notation. (R. Reagan)

67.1 Introduction

We now turn to two of the corner stones of calculus in several dimensions, namely *Gauss' theorem* and *Green's formula* , beginning with two dimensions. We shall see that these famous (and useful) results are direct consequences of the fundamental formula,

$$\int_\Omega \frac{\partial u}{\partial x_2}\,dx = \int_\Gamma u\,n_2\,ds, \tag{67.1}$$

where Ω is a domain in $\mathbb{R}^2$ with boundary Γ and $n(x) = (n_1(x), n_2(x))$ is the *outward unit normal* to Γ at $x \in \Gamma$, that is $n(x)$ is orthogonal to the tangent to Γ at x and points out of Ω and $\|n(x)\| = 1$, see Fig. 67.2. We shall see that this formula is an analog of the Fundamental Theorem

$$\int_a^b \frac{du}{dx}\,dx = u(b) - u(a), \tag{67.2}$$

stating that the integral over an interval $[a, b]$ of the derivative $\frac{du}{dx}$ of a function u is equal to the difference between the end-point values $u(b)$ and $u(a)$.

67.2 The special case with Ω a square

To see the connection between (67.1) and (67.2), we first assume that Ω is the unit square $[0,1]\times[0,1]$. In this case $n_2 = 1$ on the top Γ_1 of the square and $n_2 = -1$ on the bottom Γ_3, and $n_2 = 0$ on the vertical sides Γ_2 and Γ_4, see Fig. 67.2. Therefore,

$$\int_\Gamma u\, n_2\, ds = \int_0^1 u(x_1,1)\, dx_1 - \int_0^1 u(x_1,0)\, dx_1$$

if we parameterize Γ_1 by $s(x_1) = (x_1,1)$ and Γ_3 by $s(x_1) = (x_1,0)$. On the other hand, integrating first with respect to x_2 and then with respect to x_1 and using (67.2), we have

$$\begin{aligned}\int_\Omega \frac{\partial u}{\partial x_2}\, dx &= \int_0^1 \Big(\int_0^1 \frac{\partial u}{\partial x_2}(x_1,x_2)\, dx_2\Big)\, dx_1 = \int_0^1 \big(u(x_1,1) - u(x_1,0)\big)\, dx_1 \\ &= \int_0^1 u(x_1,1)\, dx_1 - \int_0^1 u(x_1,0)\, dx_1 = \int_\Gamma u\, n_2\, ds,\end{aligned}$$

which proves (67.1) when Ω is a square. We see that (67.1) results from using (67.2) with $\frac{du}{dx}\, dx$ replaced by $\frac{\partial u}{\partial x_2} dx_2$, followed by an integration with respect to x_1. The net result is that the integral of $\frac{\partial u}{\partial x_2} dx_2$ over Ω is replaced by a curve integral of un_2 over the boundary Γ of Ω.

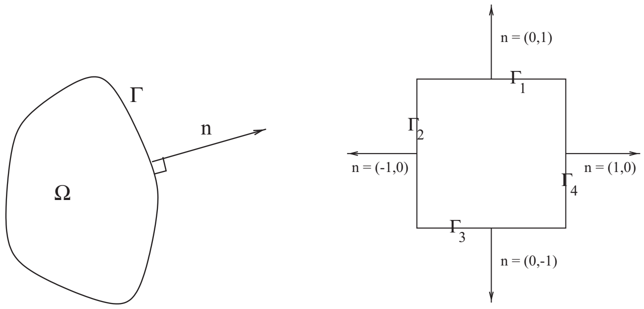


FIGURE 67.1. To the left: A domain Ω with boundary Γ and normal n. To the right: A special case.

67.3 The general case

We now consider a domain Ω bounded by two curves Γ_1 parameterized by $s_1(x_1) = (x_1, \gamma_1(x_1))$ and Γ_2 parameterized by $s_2(x_1) = (x_1, \gamma_2(x_1))$ with $a \le x_1 \le b$, and $n = (n_1, n_2)$ is the outward normal to Γ, see Fig. 67.2.

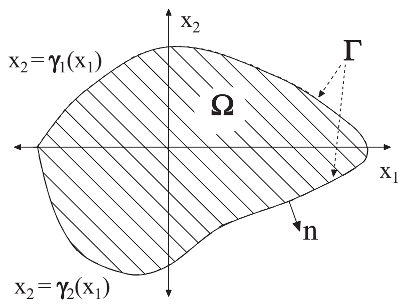


FIGURE 67.2. A domain Ω with two curves defining the boundary Γ.

The proof of (67.1) depends on the key observation that

$$\|\frac{ds_1}{dx_1}\| = \sqrt{1+(\gamma_1')^2}, \quad n_2 = \frac{1}{\sqrt{1+(\gamma_1')^2}},$$
$$\|\frac{ds_2}{dx_1}\| = \sqrt{1+(\gamma_2')^2}, \quad n_2 = -\frac{1}{\sqrt{1+(\gamma_2')^2}}.$$

Formally, $n_2 ds_1 = dx_1$ and $n_2 ds_2 = -dx_1$, see Fig. 67.3.

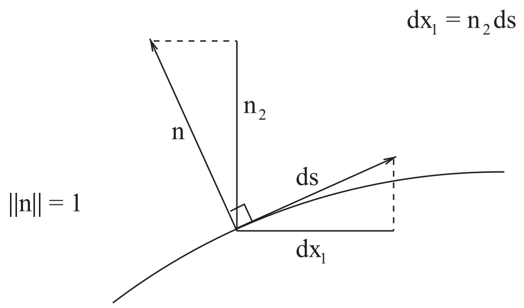


FIGURE 67.3. The key observation that $\frac{dx_1}{ds} = \frac{n_2}{1}$ by similarity.

Note that n_2 is positive on the upper boundary curve s_1 and negative on the lower boundary curve s_2. We thus have

$$\int_{\Gamma_1} un_2\,ds_1 = \int_a^b u(x_1,\gamma_1(x_1))n_2\|\frac{ds_1}{dx_1}\|\,dx_1 = \int_a^b u(x_1,\gamma_1(x_1))\,dx_1,$$
$$\int_{\Gamma_2} un_2\,ds_2 = \int_a^b u(x_1,\gamma_2(x_1))n_2\|\frac{ds_2}{dx_1}\|\,dx_1 = -\int_a^b u(x_1,\gamma_1(x_1))\,dx_1.$$

Secondly, integrating first with respect to x_2 and then with respect to x_1 and using the Fundamental Theorem, we see that

$$\int_\Omega \frac{\partial u}{\partial x_2}\, dx = \int_\Omega \frac{\partial u}{\partial x_2}\, dx_2\, dx_1 = \int_a^b \Big(\int_{\gamma_2(x_1)}^{\gamma_1(x_1)} \frac{\partial u}{\partial x_2}\, dx_2 \Big)\, dx_1$$
$$= \int_a^b u(x_1, \gamma_1(x_1))\, dx_1 - \int_a^b u(x_1, \gamma_2(x_1))\, dx_1.$$

Since

$$\int_\Gamma u\, n_2\, ds = \int_{\Gamma_1} u n_2\, ds_1 + \int_{\Gamma_2} u n_2\, ds_2,$$

the desired formula (67.1) now follows. The proof generalizes to arbitrary domains bounded by smooth curves with Lipschitz continuous tangents. We summarize in the following basic theorem:

Theorem 67.1 *If Ω is a domain in $\mathbb{R}^2$ with boundary Γ with outward unit normal (n_1, n_2) and $u : \Omega \to \mathbb{R}$, then*

$$\int_\Omega \frac{\partial u}{\partial x_i}\, dx = \int_\Gamma u\, n_i\, ds, \quad i = 1, 2. \tag{67.3}$$

Applying (67.3) to the product vw of two functions v and w, we obtain the following analog of integration by parts in two dimensions:

Theorem 67.2 (Integration by parts in 2d) *If Ω is a domain in $\mathbb{R}^2$ with boundary Γ with outward unit normal (n_1, n_2) and $v, w : \Omega \to \mathbb{R}$, then*

$$\int_\Omega \frac{\partial v}{\partial x_i} w\, dx = \int_\Gamma vw\, n_i\, ds - \int_\Omega v \frac{\partial w}{\partial x_i}\, dx, \quad i = 1, 2. \tag{67.4}$$

Applying (67.3) to the components u_i of a vector valued function $u = (u_1, u_2)$ and summing over $i = 1, 2$, we obtain the *divergence theorem*, or *Gauss' theorem*

$$\int_\Omega \nabla \cdot u\, dx = \int_\Gamma u \cdot n\, ds, \tag{67.5}$$

where $u \cdot n = u_1 n_1 + u_2 n_2$ and

$$\nabla \cdot u = \Big(\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}\Big) \cdot (u_1, u_2) = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2}.$$

Applying (67.4) with w replaced by $\frac{\partial w}{\partial x_i}$ and summing over $i = 1, 2$, we obtain *Green's formula*:

$$\int_\Omega \nabla v \cdot \nabla w\, dx = \int_\Gamma v \partial_n w\, ds - \int_\Omega v \Delta w\, dx, \tag{67.6}$$

where

$$\partial_n w = \nabla w \cdot n = \frac{\partial w}{\partial x_1} n_1 + \frac{\partial w}{\partial x_2} n_2, \tag{67.7}$$

is the *outward normal derivative* of w on Γ. We often use Green's formula in the form

$$\int_\Omega v\Delta w\,dx - \int_\Omega \Delta v\,w\,dx = \int_\Gamma v\,\partial_n w\,ds - \int_\Gamma \partial_n v\,w\,ds, \tag{67.8}$$

which results after applying (67.6) twice and using

$$\begin{aligned}\Delta w &= \text{div grad}w = \nabla\cdot\nabla w\\ &= (\frac{\partial}{\partial x_1},\frac{\partial}{\partial x_2})\cdot(\frac{\partial w}{\partial x_1},\frac{\partial w}{\partial x_2}) = \frac{\partial}{\partial x_1}(\frac{\partial w}{\partial x_1}) + \frac{\partial}{\partial x_2}(\frac{\partial w}{\partial x_2}),\end{aligned}$$

which can be written succinctly as $\Delta w = \frac{\partial^2 w}{\partial x_1^2} + \frac{\partial^2 w}{\partial x_2^2}$.

We also note the following analog of the divergence theorem:

$$\int_\Omega \nabla\times u\,dx = \int_\Gamma n\times u\,ds, \tag{67.9}$$

where $u:\Omega\to\mathbb{R}^2$ and $\nabla\times u = \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}$, and $n\times u = u_2n_1 - u_1n_2$. This is just a restatement of

$$\int_\Omega (\frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2})\,dx = \int_\Gamma (u_2n_1 - u_1n_2)\,ds \tag{67.10}$$

and therefore follows from (67.3). We further note that $\tau = (-n_2, n_1)$ is a unit tangent to Γ, since $n = (n_1, n_2)$ is a unit normal and $(-n_2, n_1)\cdot(n_1, n_2) = 0$, and $\tau = (-n_2, n_1)$ is directed in the counter-clockwise direction of Γ, see Fig. 67.4.

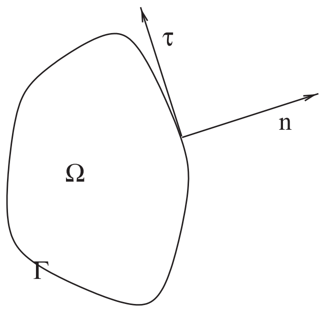


FIGURE 67.4. The unit tangent $\tau = (-n_2, n_1)$ to Γ expressed in terms of the normal $n = (n_1, n_2)$.

We often write

$$\int_\Gamma u_2n_1 - u_1n_2\,ds = \int_\Gamma u\cdot\tau\,ds = \int_\Gamma u\cdot ds,$$

interpreting ds in the last integral as the *vector* τds with the old use of ds as the element of curve length. This is consistent with the notation

$$\int_\Gamma u \cdot ds = \int_a^b u(s(t)) \cdot s'(t)\, dt,$$

where $s : [a, b] \to \mathbb{R}^2$ represents Γ, which was introduced in Chapter *Curve Integrals*. Caution: we here use "ds" with two different interpretations: as the element of curve length (a scalar), and as an element of the tangent vector (a vector).

We summarize the basic results derived in this chapter as follows:

Theorem 67.3 *If Ω is a domain in $\mathbb{R}^2$ with boundary Γ with outward unit normal (n_1, n_2), and $u : \Omega \to \mathbb{R}^2$ and $v, w : \Omega \to \mathbb{R}$, then*

$$\int_\Omega \frac{\partial v}{\partial x_i}\, dx = \int_\Gamma v\, n_i\, ds, \quad i = 1, 2, \tag{67.11}$$

$$\int_\Omega \frac{\partial v}{\partial x_i} w\, dx = \int_\Gamma vw\, n_i\, ds - \int_\Omega v \frac{\partial w}{\partial x_i}\, dx, \quad i = 1, 2, \tag{67.12}$$

$$\int_\Omega \nabla \cdot u\, dx = \int_\Gamma u \cdot n\, ds, \tag{67.13}$$

$$\int_\Omega \nabla \times u\, dx = \int_\Gamma n \times u\, ds = \int_\Gamma u \cdot ds, \tag{67.14}$$

$$\int_\Omega \nabla v \cdot \nabla w\, dx = \int_\Gamma v \partial_n w\, ds - \int_\Omega v \Delta w\, dx, \tag{67.15}$$

$$\int_\Omega v \Delta w\, dx - \int_\Omega \Delta v\, w\, dx = \int_\Gamma v\, \partial_n w\, ds - \int_\Gamma \partial_n v\, w\, ds. \tag{67.16}$$

Example 67.1. For $u(x_1, x_2) = x_1$ and $i = 1$ in (67.11), we obtain $\int_\Omega dx = \int_\Gamma x_1 n_1\, ds = \int_\Gamma x_1\, dx_2$. An interesting observation from this is that you may compute the area $\int_\Omega dx$, for example of a piece of land, simply by walking its boundary and computing $\int_\Gamma x_1\, dx_2$. The *planetometer* is a mechanical devise for computing the area of plane domains built on this principle, which has been used extensively by Surveyors.

Example 67.2. If $\nabla \times u = 0$ in the domain Ω between two curves Γ_1 and Γ_2 that both start at the point a and end at the point b, then $\int_{\Gamma_1} u \cdot ds = \int_{\Gamma_2} u \cdot ds$, where ds is the vector tangential to the curves in the direction from a to b of length equal to the element of curve length. This follows from the fact that $\int_{\Gamma_1 \cup \Gamma_2^-} u \cdot ds = \int_\Gamma u \cdot ds = 0$, by (67.14), where Γ_2^- denotes the curve Γ_2 with the direction of ds reversed. We conclude that curve integrals of a field $u = (u_1, u_2)$ with $\nabla \times u = 0$, that is of an *irrotational* field, is independent of the particular "path" of the curve from a to b. The integral of $u \cdot ds$ only depends on the two end-points a and b of the integration. Fields $u = (u_1, u_2)$ of this type are called *conservative.* As we shall see below, such fields are given by a *potential*, that is, they are the gradient field of some scalar potential $\varphi = \varphi(x)$ so $u = \nabla\varphi$.

Furthermore, $\int_\gamma u \cdot ds = \varphi(b) - \varphi(a)$ for a curve γ from a to b. For example, the field $u = (x_2, x_1)$ has $u_1(x_1, x_2) = x_2$ and $u_2(x_1, x_2) = x_1$ and thus $\nabla u = \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2} = 1 - 1 = 0$. We find easily that $u = \nabla\varphi$ for $\varphi(x) = x_1 x_2$, and the integral of $u \cdot ds$ from a point $a = (a_1, a_2)$ to $b = (b_1, b_2)$ is given by $b_1 b_2 - a_1 a_2$.

Chapter 67 Problems

67.1. Derive (67.4), (67.5), (67.6) and (67.8) from (67.3).

67.2. (a) Explain why (67.1) is valid also for a domain like $\{(x_1, x_2) : x_1 \leq |x_2|, x_1^2 + x_2^2 \leq 1\}$. (b) Verify by direct computation of $\int_\Omega \frac{\partial u}{\partial x_2}\, dx$ and $\int_\Gamma u\, n_2\, ds$ that (67.1) is valid for $u = r^{1/4}\sin(v/4)$ and $\Omega = \{(r\cos(v), r\sin(v)) : 0 < r < 1, 0 < v < 2\pi\}$, where $r = \sqrt{x_1^2 + x_2^2}$ and $v = \operatorname{arccot}(x_1/x_2)$ for $x_2 > 0$, $v = \operatorname{arccot}(x_1/x_2) + \pi$ for $x_2 < 0$ are the usual polar coordinates. Recall that by the chain rule you may express $\frac{\partial u}{\partial x_2}$ in terms of $\frac{\partial u}{\partial r}$ and $\frac{\partial u}{\partial v}$ if you like.

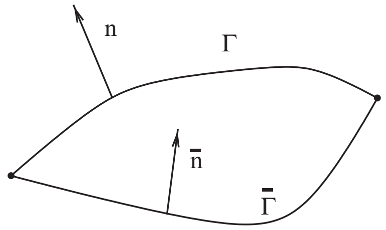
67.3. Assume $u = (u_1, u_2)$ is divergence free in Ω with boundary Γ. What can be said about (a) $\int_\Gamma u \cdot n\, ds$, (b) $u(x) \cdot n(x)$ for points x on Γ.

67.4. Assume $\int_\Gamma u \cdot n\, ds = 0$, where Γ is the boundary of a domain Ω with exterior unit normal n. What can be said about $\nabla \cdot u$ in Ω? (Before you give a too definite answer you may want to consider for example the case $u = (x_1^2, x_2^2)$ with Ω the unit disc.) Assume $\int_\gamma u \cdot n\, ds = 0$ for all closed curves γ in Ω, and the derivatives of u_i are Lipschitz. What can then be said about u in Ω?

67.5. Consider a "deformation" of $\mathbb{R}^2$ where the points $x = (x_1, x_2)$ are displaced to new positions $x + u(x)$, $u = (u_1, u_2)$, $u_i = u_i(x)$. We call $u(x)$ the displacement field and the Jacobian $u'(x)$ of $u(x)$ the deformation tensor (matrix). Consider for simplicity the case $u_i(x) = a_i x_i, i = 1, 2$, and assume the displacement is "area preserving", corresponding to "incompressibility" of the deformed material. Show that for small deformations, one has div $u \approx 0$. Hint: Consider $x \to x + u(x)$ as a change of variables and use an established fact about the Jacobian of area preserving maps.

67.6. Consider the vector field $u(x) = x/\|x\|^2$. Let Ω be the disc $\{x \in \mathbb{R}^2 : \|x - a\| \leq 1\}$, and Γ its boundary with exterior unit normal n. Compute $\int_\Gamma u \cdot n\, ds$ for $a = (2, 0)$ and $a = 0$. Do the results conform with the Divergence theorem? Make an "arrow plot" of u in the (x_1, x_2)-plane. Can you see a connection the eruption of a volcano? Does the Divergence theorem apply in the case $a = (0, 0)$?

67.7. Show that if $\nabla \cdot u = 0$ and Γ and $\bar{\Gamma}$ are curves with normals n and $\bar{n}$ as in Fig. 67.7, then $\int_\Gamma u \cdot n\, ds = \int_{\bar{\Gamma}} u \cdot \bar{n}\, ds$.



67.8. For $u = (\frac{2x_1x_2}{1+x_2^2}, -\log(1+x_2^2))$ and Γ the curve (a) $\{(x_1,x_2) : x_1^2 + x_2^2 = 1, x_i \geq 0, i = 1,2\}$ (b) $\{(x_1,x_2) : x_1 = 2-(x_2-1)^2, x_1 \geq 1\}$, compute $\int_\Gamma u \cdot n\, ds$. Hint: Close the curves and use the Divergence theorem.

67.9. Show that the field $u = e^{x_1x_2}(1 + x_1x_2, x_1^2)$ is irrotational, and find a potential φ such that $u = \nabla\varphi$.

67.10. Evaluate the integrals in (67.16) for w a solution to the *differential equation* $-\Delta w = f$ in $\Omega = \mathbb{R}^2$ and $v = -\frac{1}{2\pi}\log(x - \bar{x})$, assuming w and $\partial_n w$ vanish for $\|x\|$ large. Show that this gives a formula for $w(\bar{x})$ in terms of f and v. Hint: Take $\Omega = \{x \in \mathbb{R}^2 : \|x - \bar{x}\| > \epsilon\}$ and let ϵ tend to zero.

67.11. Let w be the solution to $-\Delta w = f$ in the upper half plane $x_2 > 0$, $-\frac{\partial w}{\partial x_2} = g$ for $x_2 = 0$, and assume w and ∇w vanish for $\|x\|$ large. Show that for $\bar{x} = (\bar{x}_1, 0)$ on $\Gamma = \{(x_1,x_2) : x_2 = 0\}$, one has $\frac{1}{2}w(\bar{x}) = \int_{\{x:x_2>0\}} vf\, dx + \int_{\{x:x_2=0\}} vg\, ds$, where $v = -\frac{1}{2\pi}\log(x - \bar{x})$. Hint: Take $\Omega = \{x \in \mathbb{R}^2 : x_2 > 0, \|x - \bar{x}\| > \epsilon\}$ in (67.16), and let ϵ tend to zero.

67.12. Show that for *harmonic functions* v and w, that is with $\Delta v = 0$ and $\Delta w = 0$, one has $\int_\Gamma \partial_n v w\, ds = \int_\Gamma v \partial_n w\, ds$ for a closed curve Γ.

67.13. Find the area of the domain enclosed by the curve $\Gamma = \{(r\cos(v), r\sin(v)) : r = 2 + \sin(v), 0 \leq v < 2\pi\}$. Hint: Integrals of the form $\int \sin^4(v)\, dv$ and $\int \cos^4(v)\, dv$ may be computed using integration by parts, as follows:

$$\begin{aligned}
I = \int \cos^4(v)\, dv &= \int (1 - \sin^2(v))\cos(v) \cdot \cos(v)\, dv = \{\text{int. by parts}\} \\
&= (\sin(v) - \frac{1}{3}\sin^3(v)) \cdot \cos(v) - \int (\sin(v) - \frac{1}{3}\sin^3(v))(-\sin(v))\, dv \\
&= (\sin(v) - \frac{1}{3}\sin^3(v)) \cdot \cos(v) + \int \sin^2(v)\, dv - \frac{1}{3}\int (1 - \cos^2(v))^2\, dv \\
&= (\sin(v) - \frac{1}{3}\sin^3(v))\cos(v) + \int \sin^2(v)\, dv - \frac{1}{3}\int (1 - 2\cos^2(v))\, dv - \frac{1}{3}I,
\end{aligned}$$

from which I can be computed.

68

Gauss' theorem and Green's formula in $\mathbb{R}^3$

Of those who with me have written something about these matters, either I alone am mad, or I alone am not mad. No third option can be maintained, unless (as perchance it may seem to some) we are all mad. (Hobbes to Wallis)

If he is mad, he is not likely to be convinced by reason; on the other hand, if we be mad, we are in no position to attempt it. (Wallis to Hobbes)

68.1 Introduction

We now extend the results of the previous chapter to three dimensions. The basic result is the following analog of (67.1): If Ω is a domain in $\mathbb{R}^3$ with boundary Γ, then

$$\int_\Omega \frac{\partial u}{\partial x_3}\,dx = \int_\Gamma u\,n_3\,ds, \tag{68.1}$$

where (n_1, n_2, n_3) is the outward normal to Γ. To prove this, we assume that Γ is composed of the two surfaces Γ_1 given by $s_1(x_1, x_2) = (x_1, x_2, \gamma_1(x_1, x_2))$ and Γ_2 given by $s_2(x_1, x_2) = (x_1, x_2, \gamma_2(x_1, x_2))$, where $(x_1, x_2) \in \omega$ and the parameter domain ω is a domain in $\mathbb{R}^2$, and we assume that $\Omega = \{x \in \mathbb{R}^3 : (x_1, x_2) \in \omega, \gamma_2(x_1, x_2) < x_3 < \gamma_1(x_1, x_2)\}$, see Fig. 68.1. We have $s'_{i,1} \times s'_{i,2} = (1, 0, \gamma_{i,1}) \times (0, 1, \gamma_{i,2})$ for $i = 1, 2$, where $\gamma_{i,j} = \frac{\partial \gamma_i}{\partial x_j}$, and thus on Γ_1

$$\|s'_{1,1} \times s'_{1,2}\| = \sqrt{1 + (\gamma'_{1,1})^2 + (\gamma'_{1,2})^2}, \quad n_3 = \frac{1}{\sqrt{1 + (\gamma'_{1,1})^2 + (\gamma'_{1,2})^2}}$$

and on Γ_2

$$\|s'_{2,1} \times s'_{2,2}\| = \sqrt{1 + (\gamma'_{2,1})^2 + (\gamma'_{2,2})^2}, \quad n_3 = -\frac{1}{\sqrt{1 + (\gamma'_{2,1})^2 + (\gamma'_{2,2})^2}}.$$

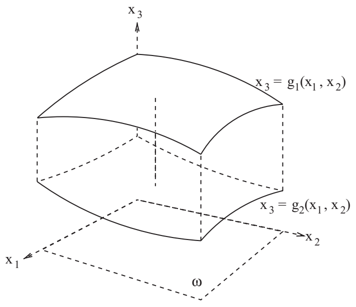


FIGURE 68.1. A domain Ω bounded by two graphs Γ_1 and Γ_2.

Integrating first with respect to x_3 and using the Fundamental Theorem we get

$$\begin{aligned}
&\int_\Omega \frac{\partial u}{\partial x_3}\, dx = \int_\Omega \frac{\partial u}{\partial x_3}\, dx_3 dx_1 dx_2 \\
&= \int_\omega u(x_1, x_2, \gamma_1(x_1, x_2))\, dx_1 dx_2 - \int_\omega u(x_1, x_2, \gamma_2(x_1, x_2))\, dx_1 dx_2 \\
&= \int_\omega u\, n_3 \|s'_{1,1} \times s'_{1,2}\| dx_1 dx_2 + \int_\omega u\, n_3 \|s'_{2,1} \times s'_{2,2}\| dx_1 dx_2 \\
&= \int_{\Gamma_1} u\, n_3\, ds + \int_{\Gamma_2} u\, n_3\, ds = \int_\Gamma u\, n_3\, ds,
\end{aligned}$$

which proves (68.1). Note that $n_3 = 0$ on the "vertical" parts of Γ! This result generalizes to

$$\int_\Omega \frac{\partial u}{\partial x_i}\, dx = \int_\Gamma u\, n_i\, ds, \quad i = 1, 2, 3, \tag{68.2}$$

for a general domain Ω in $\mathbb{R}^3$ with boundary Γ with outward unit normal (n_1, n_2, n_3).

Applying (68.2) to the product vw of two functions v and w, we obtain the analog of integration by parts in three dimensions,

$$\int_\Omega \frac{\partial v}{\partial x_i} w\, dx = \int_\Gamma vw\, n_i\, ds - \int_\Omega v \frac{\partial w}{\partial x_i}\, dx, \quad i = 1, 2, 3. \tag{68.3}$$

Applying (68.3) to the components u_i of a vector valued function $u = (u_1, u_2, u_3)$ with $w = 1$ and summing over i, we obtain the *divergence theorem*, or *Gauss' theorem* in three dimensions

$$\int_\Omega \nabla \cdot u\, dx = \int_\Gamma u \cdot n\, ds, \tag{68.4}$$

where $\nabla \cdot u = (\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_3}) \cdot (u_1, u_2, u_3) = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3} = \sum_{i=1}^3 \frac{\partial u_i}{\partial x_i}$, and $u \cdot n = u_1 n_1 + u_2 n_2 + u_3 n_3$ is the component of u in the direction of the normal n. If u represents a *flux* of some quantity, like heat flux or water flux, then $u(x) \cdot n(x)$ at a point $x \in \Gamma$ represents the flux through Γ (out of Ω), or *normal flux*, and thus

$$\int_\Gamma u \cdot n\, ds$$

represents the *total flux* through Γ.

We also directly obtain the following analog of Gauss' theorem for a function $u : \Omega \to \mathbb{R}^3$:

$$\int_\Omega \nabla \times u\, dx = \int_\Gamma n \times u\, ds, \tag{68.5}$$

which is now a vector equation!!

Another consequence of (68.3) is *Green's formula*:

$$\int_\Omega \nabla v \cdot \nabla w\, dx = \int_\Gamma v \partial_n w\, ds - \int_\Omega v \Delta w\, dx, \tag{68.6}$$

where $\partial_n v = \nabla v \cdot n = \frac{\partial v}{\partial x_1} n_1 + \frac{\partial v}{\partial x_2} n_2 + \frac{\partial v}{\partial x_3} n_3$ is the outward normal derivative of v on Γ, and now $\Delta w = \frac{\partial^2 w}{\partial x_1^2} + \frac{\partial^2 w}{\partial x_2^2} + \frac{\partial^2 w}{\partial x_3^2}$. We often use Green's formula in the form

$$\int_\Omega v \Delta w\, dx - \int_\Omega \Delta v\, w\, dx = \int_\Gamma v\, \partial_n w\, ds - \int_\Gamma \partial_n v\, w\, ds, \tag{68.7}$$

which results after applying (68.6) twice.

We summarize the basic results derived in this chapter as follows:

Theorem 68.1 *If Ω is a domain in $\mathbb{R}^3$ with boundary Γ with outward unit normal $n = (n_1, n_2, n_3)$, and $u : \Omega \to \mathbb{R}^3$ and $v, w : \Omega \to \mathbb{R}$, then*

$$\int_\Omega \frac{\partial v}{\partial x_i}\, dx = \int_\Gamma v\, n_i\, ds, \quad i = 1, 2. \tag{68.8}$$

$$\int_\Omega \frac{\partial v}{\partial x_i} w\,dx = \int_\Gamma vw\,n_i\,ds - \int_\Omega v\frac{\partial w}{\partial x_i}\,dx, \quad i=1,2. \tag{68.9}$$

$$\int_\Omega \nabla\cdot u\,dx = \int_\Gamma u\cdot n\,ds \tag{68.10}$$

$$\int_\Omega \nabla\times u\,dx = \int_\Gamma n\times u\,ds, \tag{68.11}$$

$$\int_\Omega \nabla v\cdot\nabla w\,dx = \int_\Gamma v\partial_n w\,ds - \int_\Omega v\Delta w\,dx, \tag{68.12}$$

$$\int_\Omega v\Delta w\,dx - \int_\Omega \Delta v\,w\,dx = \int_\Gamma v\,\partial_n w\,ds - \int_\Gamma \partial_n v\,w\,ds. \tag{68.13}$$

Example 68.1. We compute the total flow of the vector field $u(x) = (x_1+x_2^5, x_2+x_3x_1, x_3+x_1x_2)$ out of the boundary S of the unit ball $B=\{x\in\mathbb{R}^3 : \|x\|=1\}$, that is the integral,

$$\int_S u\cdot n\,ds = \int_S ((x_1+x_2^5)x_1 + (x_2+x_3x_1)x_2 + (x_3+x_1x_2)x_3)\,ds, \tag{68.14}$$

where we used that the outward unit normal n to S at $x\in S$ is given by $n(x)=x$. Since div $u(x)=3$ for $x\in\mathbb{R}^3$, we have by Gauss's theorem

$$\int_S u\cdot n\,ds = \int_B 3\,dx = 3V(B) = 4\pi,$$

which gives a quick way of computing the quite difficult integral (68.14).

Chapter 68 Problems

68.1. Write out and verify (68.5) from (68.2).

68.2. (a) Prove Green's formula (68.6) using (68.3). (b) Prove (68.13).

68.3. Compute the integral $\int_\Gamma \frac{x}{\|x\|^3}\cdot n\,ds$, where $\Gamma=\{x\in\mathbb{R}^3 : x_1^2+x_2^2+(x_3-ja)^2=a^2\}$ for $a>0$ and $j=0,1,2$, respectively, where n is the exterior unit normal to Γ. Interprete the results.

68.4. Compute the integral $\int_\Gamma \frac{1}{x_1^2+x_2^2}\frac{(-x_2,x_1)}{x_1^2+x_2^2}\times ds$, where $\Gamma=\{x\in\mathbb{R}^3 : x_1^2+x_2^2+x_3=1\}$.

68.5. Let Γ be the unit sphere in $\mathbb{R}^3$ with exterior unit normal n and compute the following integrals:

1. $\int_\Gamma x\cdot n\,ds$,

2. $\int_\Gamma x \times n\, ds$,
3. $\int_\Gamma \frac{1}{\|x\|^2} \frac{x}{\|x\|} \cdot n\, ds$,
4. $\int_\Gamma \frac{1}{\|x\|^2} \frac{x}{\|x\|} \times n\, ds$.

68.6. Verify that for a radial field $F(x) = \|x\|^\alpha \frac{x}{\|x\|}$ one has div $F = (\alpha + 2)\|x\|$.

68.7. What is the smallest possible value of the integral $\int_\Gamma F \cdot n\, ds$, where $F(x) = (x_1 x_2^2 - 4x_1 x_2, 4x_2 x_3^2 + 8x_2 x_3 + 5x_2, x_1^2 x_3 - 2x_1 x_3)$ and Γ is a closed surface in $\mathbb{R}^3$, and n its exterior unit normal? Hint: Enclose all the "sinks" of F, that is, consider the domain where div $F \leq 0$.

68.8. Compute the surface integral $\int_\Gamma F \cdot n\, ds$, where

$$F(x) = (x_2^2, x_1 x_2 (\cos(x_1))^2 + x_1 x_2^3 + \exp(\cos(x_1 x_3^2)), x_1 x_3 (\sin(x_1))^2 - 3x_1 x_2^2 x_3),$$

and Γ is the part of the sphere $\|x\| = 2$ with positive x_3-coordinate, and n its normal with also positive x_3-component. Hint: The function F is chosen *seemingly* difficult only to confuse you.

68.9. Let $\{x_1, x_2, ..., x_N\}$ be a set of points in $\mathbb{R}^n$, and let

$$F(x) = \sum_{j=1}^{N} \frac{1}{4\pi \|x - x_j\|^2} \frac{x - x_j}{\|x - x_j\|}.$$

Compute the surface integral $\int_\Gamma F \cdot n\, ds$ for any closed surface Γ containing $k \leq N$ of the points $\{x_1, x_2, ..., x_N\}$, with n the exterior unit normal as usual.

68.10. Show, as you did in Chapter Newton's Nightmare, that the gravitation from a sphere is the same as if all the mass of the sphere was concentrated to its center, but now using Gauss' theorem to make things easier. Use as starting point that the divergence of the gravitational field is (proportional to) the density, i.e.

$$\nabla \cdot F = \rho / c$$

for some constant c, and assume spherical symmetry, i.e. the direction of the gravitational field is in the radial direction from the center of the sphere.

68.11. Show that if $-\Delta u = f$ in Ω, then for any function v that is zero on Γ, the boundary of Ω, one has

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega f v.$$

Also prove that if $\partial_n u = g$ on Γ, then for all functions v,

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega f v + \int_\Gamma g v\, ds.$$

69
Stokes' theorem

> I too feel that I have been thinking too much of late, but in a different way, my head running on divergent series, the discontinuity of arbitrary constants, ... I often thought that you would do me good by keeping me from being too engrossed by those things. (Stokes asking Mary Susanna Robinson to marry him 1857)

69.1 Introduction

Stokes' theorem states that if $u : \mathbb{R}^3 \to \mathbb{R}^3$ is differentiable, then

$$\int_S (\nabla \times u) \cdot n \, ds = \int_\Gamma u \cdot ds\,, \tag{69.1}$$

where S is a surface in $\mathbb{R}^3$ bounded by a closed curve Γ, n is a unit normal to S, and Γ is oriented in a clockwise direction following the positive direction of the normal n, see Fig. 69.1. The integral

$$\int_\Gamma u \cdot ds$$

is called the *circulation* of u around Γ. The integral

$$\int_S (\nabla \times u) \cdot n \, ds$$

is the *total flow of the rotation* $\nabla \times u$ *across the surface* S. Stokes' theorem states that the total flow of $\nabla \times u$ across S is equal to the circulation of u around the boundary Γ of S.

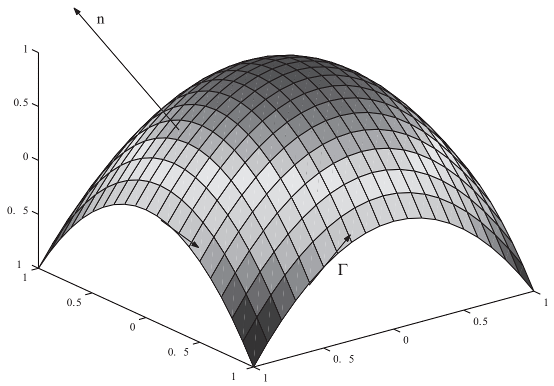


FIGURE 69.1. A Stokes surface S with boundary curve Γ.

Stokes (1819-1903), see Fig. 69.2, an Irish mathematician/physicist and professor in Cambridge 1849, gave basic contributions to the theory of viscous fluid flow modeled by the Navier-Stokes equations.

FIGURE 69.2. Stokes

69.2 The special case of a surface in a plane

We start by considering the special case of a plane surface $\bar{S}$ in the plane $\{x \in \mathbb{R}^3 : x_3 = 0\}$ with normal $\bar{n} = (0, 0, 1)$ and with boundary Γ, see

Fig. 69.3. In this case, Stokes' theorem takes the form

$$\int_{\bar{S}} (\nabla \times u) \cdot \bar{n}\, ds = \int_{\bar{S}} \left(\frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2} \right) dx_1\, dx_2$$
$$= \int_\Gamma u \cdot ds = \int_\Gamma (u_2 n_1 - u_1 n_2)\, ds. \quad (69.2)$$

By identifying the plane $\{x_3 = 0\}$ with $\mathbb{R}^2$, this is (67.10) and is a direct consequence of (67.3). This result is often referred to as *Green's formula in two dimensions*. We have thus proved Stokes' theorem in the case of a plane surface S in the plane $\{x_3 = 0\}$.

Note that the unit tangent direction is given by $\tau = (-\tilde{n}_2, \tilde{n}_1)$, where $\tilde{n} = (\tilde{n}_1, \tilde{n}_2)$ is the outward normal direction to Γ in the plane $\{x : x_3 = 0\}$ with a counter clockwise orientation when viewed from the top of the normal $\bar{n} = (0, 0, 1)$ of $\bar{S}$. The orientation is consistent with the specification that τ should be oriented clockwise when following the direction of the normal to $\bar{S}$.

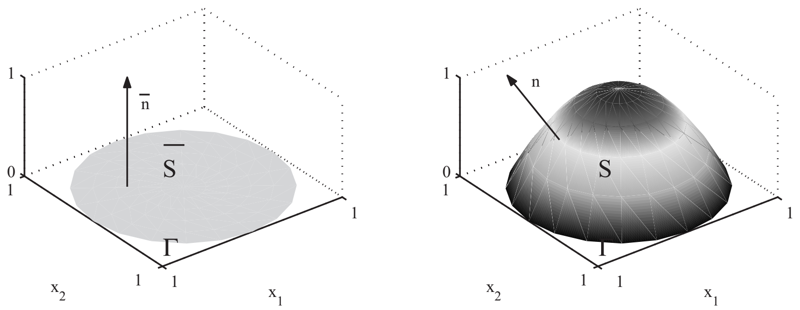


FIGURE 69.3. Two special cases: A plane surface $\bar{S}$ with normal $\bar{n}$ and boundary curve Γ, and a curved surface S with normal n and a plane boundary curve Γ.

EXAMPLE 69.1. Let $S = \{x \in \mathbb{R}^3 : \|x\| \leq 1, x_3 = 0\}$ be the unit disc in the plane $\{x_3 = 0\}$ bounded by the curve Γ parameterized by $s(t) = (\cos(t), \sin(t), 0)$, $0 \leq t \leq 2\pi$. Choose $n = (0, 0, 1)$ and let $u(x) = (-x_2, x_1, 0)$ so that $\nabla \times u(x) = (0, 0, 2)$. We compute

$$\int_S (\nabla \times u)\, ds = 2\pi, \quad \int_\Gamma u \cdot ds = \int_0^{2\pi} (\cos^2(t) + \sin^2(t)) dt = 2\pi,$$

in accordance with Stokes' theorem.

EXAMPLE 69.2. Ampere's law states the $\nabla \times H = J$, where H is the magnetic field and J the electric current. Stokes' theorem states that the circulation of H around a closed curve Γ bounding a surface S is equal to the total current through the surface S. Stokes' theorem is thus one of the corner-stones of electromagnetic field theory.

69.3 Generalization to an arbitrary plane surface

We shall now verify that both the left and right hand side of Stokes' equality

$$\int_S (\nabla \times u) \cdot n \, ds = \int_\Gamma u \cdot ds\,,$$

are invariant under orthogonal coordinate transformations. We thus obtain a proof of Stokes' theorem for a given plane surface S through the origin, by choosing coordinates so that S lies in the plane $\{x_3 = 0\}$, and using the proof of the previous section. The case of a surface S not passing through the origin is reduced to the previous case by a simple translation of the origin of the coordinate system.

To prove the invariance, let $x = Q\bar{x}$ be an orthogonal coordinate transformation with Q an orthogonal 3×3 matrix from a set of coordinates $\bar{x}$ to x. The dependent vector variable u also transforms as $u = Q\bar{u}$, where u are the components in x-coordinates and $\bar{u}$ the coordinates of the same quantity in $\bar{x}$-coordinates. We have a similar relation between the elements of integration $ds = s'(t)dt$ and $d\bar{s} = \bar{s}'(t)dt$ in the different coordinates since $s'(t) = Q\bar{s}'(t)$, that is $ds = Qd\bar{s}$. Therefore,

$$\int_\Gamma u \cdot ds = \int_\Gamma Q\bar{u} \cdot Qd\bar{s} = \int_\Gamma Q^\top Q\bar{u} \cdot d\bar{s} = \int_\Gamma \bar{u} \cdot d\bar{s},$$

and the invariance of the right hand side of (69.1) follows.

To prove the invariance of the left hand side of (69.1), we use the Chain rule to obtain the following relation between the gradient ∇ with respect to x and the gradient $\bar{\nabla}$ with respect to $\bar{x}$,

$$\nabla = Q\bar{\nabla}.$$

A direct computation shows that

$$(\nabla \times u) \cdot n = (Q\bar{\nabla} \times Q\bar{u}) \cdot Q\bar{n} = (\bar{\nabla} \times \bar{u}) \cdot \bar{n}, \tag{69.3}$$

which proves the invariance since $d\bar{x} = dx$. Note that (69.3) is analogous to the the relation

$$(Qa \times Qb) \cdot Qc = (a \times b) \cdot c$$

for $a, b, c \in \mathbb{R}^3$. This expresses the invariance of the volume spanned by three vectors a, b and c under orthogonal coordinate transformations.

69.4 Generalization to a surface bounded by a plane curve

Suppose that S is a surface bounded by a curve Γ contained in the plane $\{x_3 = 0\}$, see Fig. 69.3. We do not assume that S is contained in $\{x_3 = 0\}$.

Let $\bar{S}$ be the surface in the plane $\{x_3 = 0\}$ with the boundary Γ and let Ω be the volume bounded by the surface S and the plane surface $\bar{S}$. Since $\nabla \cdot (\nabla \times u) = 0$, the Divergence theorem implies

$$0 = \int_\Omega \nabla \cdot (\nabla \times u)\, dx = \int_S \nabla \times u \cdot n\, ds + \int_{\bar{S}} \nabla \times u \cdot n\, ds, \qquad (69.4)$$

where n is the outward unit normal to the boundary $\partial\Omega$ of Ω. If n is a normal to S and $\bar{n} = -n$ is a normal to $\bar{S}$, then (69.4) implies

$$\int_S \nabla \times u \cdot n\, ds = \int_{\bar{S}} \nabla \times u \cdot \bar{n}\, ds.$$

Applying Stokes' theorem to $\bar{S}$, we obtain

$$\int_S \nabla \times u \cdot n\, ds = \int_{\bar{S}} \nabla \times u \cdot \bar{n}\, ds = \int_\Gamma u \cdot ds,$$

which proves Stokes' theorem for the surface S bounded by the plane curve Γ.

A proof of Stokes' theorem for the case of a general curve is outlined in Problem 69.1. We now summarize:

Theorem 69.1 (Stokes' theorem). *If S is a surface in $\mathbb{R}^3$ with unit normal n, and Γ is the boundary of S oriented clockwise following the direction of n, then*

$$\int_S (\nabla \times u) \cdot n\, ds = \int_\Gamma u \cdot ds.$$

We state the following important direct consequence of Stokes' theorem:

Theorem 69.2 *If $u : \Omega \to \mathbb{R}^3$ with Ω a domain in $\mathbb{R}^3$ is a differentiable vector field such that*

$$\int_\Gamma u \cdot ds = 0$$

for all closed curves Γ in Ω, then $\nabla \times u = 0$ in Ω.

Chapter 69 Problems

69.1. Prove Stokes theorem for a curve Γ given by $s(t) = (x_1(t), x_2(t), f(x_1(t), x_2(t)))$, $t \in [a, b]$, where $f : \mathbb{R}^2 \to \mathbb{R}$, bounding a surface $\Omega = \{x \in \mathbb{R}^3 : x_3 - f(x_1, x_2) = 0\}$ in $\mathbb{R}^3$. Hint: The projection of Γ on the x_1x_2-plane is the curve $\tilde{\Gamma}$ represented by

$\tilde{s}(t) = (x_1(t), x_2(t), 0)$ which bounds the domain $\tilde{\Omega}$ in the x_1x_2-plane. Show that, writing $u_i = u_i(x_1, x_2, f(x_1, x_2))$,

$$\int_\Gamma u \cdot ds = \int_a^b \big(u_1 x_1' + u_2 x_2' + u_3(\frac{\partial f}{\partial x_1} x_1' + \frac{\partial f}{\partial x_2} x_2')\big)\, dt$$
$$= \int_a^b \big((u_1 + u_3 \frac{\partial f}{\partial x_1}) x_1' + (u_2 + u_3 \frac{\partial f}{\partial x_2}) x_2'\big)\, dt = \int_{\tilde{\Gamma}} \big(u_1 + u_3 \frac{\partial f}{\partial x_1}, u_2 + u_3 \frac{\partial f}{\partial x_2}\big) \cdot ds = I.$$

Then use the Stokes theorem for a plane curve established above, to show that

$$I = \int_{\tilde{\Omega}} \big(\frac{\partial}{\partial x_1}(u_2 + u_3 \frac{\partial f}{\partial x_2}) - \frac{\partial}{\partial x_2}(u_1 + u_3 \frac{\partial f}{\partial x_1})\big)\, dx,$$

and prove by performing the differentiations and direct computation that

$$I = \int_\Omega (\nabla \times u) \cdot n\, ds,$$

where $n\, ds = (-\frac{\partial f}{\partial x_1}, -\frac{\partial f}{\partial x_2}, 1)\, dx$.

69.2. Give a proof of the equality $\int_\Omega \nabla \times u\, dx = \int_\Gamma n \times u\, ds$, where Ω is a subset of $\mathbb{R}^3$ with boundary Γ with outward unit normal n, by applying the divergence theorem to $u \times a$ with a an arbitrary constant vector.

69.3. Study the relation between Green's formula (67.9), in the form (67.10), and the divergence theorem applied to the two-dimensional domain S with boundary Γ:

$$\int_S \left(\frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2}\right) dx_1\, dx_2 = \int_\Gamma (v_1 n_1 + v_1 n_2)\, ds\,,$$

with the identification $(u_1, u_2) = (-v_2, v_1)$ corresponding to counter clockwise rotation of the vector (v_1, v_2) by $\pi/2$. Explain how the clockwise direction in Stokes' theorem becomes a counter clockwise direction in (67.9).

69.4. Compute the integral

$$\int_\Gamma (x_1^2 x_2, -x_1^3)/\|x\|^4 \cdot ds,$$

where Γ is the curve in the x_1x_2-plane from $(1,0)$ to $(0,1)$ defined by $(x_1(t), x_2(t)) = (\cos(t)^{15}, \sin(t)^{17})$, $0 \le t \le \pi/2$.

69.5. Compute the integral

$$\int_\Gamma \frac{1}{x_1^2 + x_2^2} \frac{(-x_2, x_1, x_3)}{\|x\|} \cdot ds,$$

where Γ is a curve traversing the unit circle in the x_1x_2-plane five times counterclockwise, then two times clockwise, and then again four times counterclockwise, as viewed from the positive x_3-axis.

69.6. Use Stokes' theorem to prove that

$$\int_\Gamma v\, ds = \int_S n \times \nabla v\, ds,$$

where S is a surface in $\mathbb{R}^3$ bounded by the closed curve Γ. Hint: Use Stokes' theorem with $u = va$ and a is a arbitrary vector in $\mathbb{R}^3$.

69.7. Verify by direct computation Stokes' theorem for (a) S the hemisphere $\{x \in \mathbb{R}^3 : \|x\| = 1,\, x_3 \geq 0\}$ and $u = (x_2, 2x_3, 3x_1)$, (b) $S = \{x \in \mathbb{R}^3 : x_3 = 1 - x_1^2 - x_2^2,\, x_3 \geq 0\}$.

69.8. (a) Let Ω be a domain in $\mathbb{R}^2$ with boundary Γ. Show that the area $A(\Omega)$ is given by the formula

$$A(\Omega) = \frac{1}{2}\int_\Gamma u \cdot ds,$$

where $u(x) = (-x_2, x_1)$ and Γ is oriented counter-clockwise. Use this result to show that the area bounded by the ellipse $x = (a\cos(t), b\sin(t))$, $0 \leq t \leq 2\pi$, with half-axes a and b, is equal to πab. (b) Try to design a mechanical instrument for measuring the area of a domain in $\mathbb{R}^2$ (planimeter).

70
Potential fields

> He is a rather tall, lanky-looking man, with moustache and beard about to turn grey with a somewhat harsh voice and rather deaf. He was unwashed, with his cup of coffee and cigar. One of his failings is forgetting time, he pulls his watch out, finds it past three, and runs out without even finishing the sentence. (Thomas Hirst about Dirichlet 1850)

70.1 Introduction

We know from Chapter Curve integrals that potential force fields play an important role in mechanics. Let $u : \Omega \to \mathbb{R}^3$ be a given vector function, where Ω is a domain in $\mathbb{R}^3$. How can we check if $u(x)$ is a *potential field*, that is, if there is a scalar function or scalar *potential*, φ such that

$$u(x) = \nabla\varphi(x) \quad \text{for } x \in \Omega? \tag{70.1}$$

We recall that if $u = \nabla\varphi$ is a potential field and Γ is a curve parameterized by $s : [0, 1] \to \mathbb{R}^3$ from $a = s(0)$ to $b = s(1)$, then the work of u along Γ is given by

$$\int_\Gamma u \cdot ds = \int_0^1 \nabla\varphi(s(t)) \cdot s'(t)\, dt = \int_0^1 \frac{d\varphi(s(t))}{dt}\, dt = \varphi(b) - \varphi(a).$$

In particular, the work is the same along all curves from a to b, and if the curve is closed with $\varphi(1) = \varphi(0)$ then the work performed when moving around the curve is zero. A field with the property that the work along a closed curve is zero is referred to as a *conservative field.* A potential field is thus a conservative field.

A basic example of a gradient field is the gravitational field of a mass m at the origin,

$$u(x) = -m\frac{x}{\|x\|^3} = \nabla(\frac{m}{\|x\|}),$$

normalizing units so the gravitational constant is one. The electrical field of a charge m at the origin has the same form. In that case, the potential $\varphi(x) = m/\|x\|$ represents *potential energy* (gravitational or electrical), and curve integrals $\int_\Gamma u \cdot ds = \varphi(b) - \varphi(a)$ represents the work performed by a unit mass or charge when moved from a to b along Γ.

70.2 An irrotational field is a potential field

We saw earlier that a potential field $u = \nabla\varphi$ is *irrotational*, that is $\nabla \times u = \nabla \times (\nabla\varphi) = 0$. This follows from a direct computation using $\frac{\partial^2\varphi}{\partial x_i \partial x_j} = \frac{\partial^2\varphi}{\partial x_j \partial x_i}$. In other words, $\nabla \times u = 0$ in Ω is a *necessary condition* for u to be a potential field in Ω.

We shall now prove that, the condition $\nabla \times u = 0$ in Ω is a *sufficient condition* for u to be a potential field in Ω, under the assumption that Ω is convex. We recall that Ω is *convex* if for any two points x and $\bar{x}$ in Ω, the entire line segment $\bar{x} + t(x - \bar{x}), 0 \leq t \leq 1$ between $\bar{x}$ and x, is also in Ω, see Fig. 70.1. Convexity implies in particular that Ω has "no holes". We

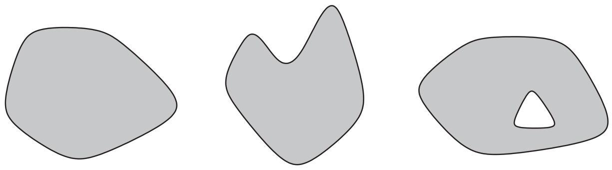


FIGURE 70.1. One convex and two non-convex domains.

thus conclude that u is a potential field in a convex domain Ω if and only u is irrotational in Ω. In other words, $u = \nabla\varphi$ in Ω for some potential φ if and only if $\nabla \times u = 0$ in Ω.

We carry out the proof by *constructing* a potential φ such that $\nabla\varphi = u$ for a given irrotational field $u(x)$ in the convex domain Ω. For the construction, we choose a fixed point $\bar{x}$ in Ω. For each point x, we let Γ_x be a curve in Ω connecting $\bar{x}$ to x and we define

$$\varphi(x) = \int_{\Gamma_x} u \cdot ds. \tag{70.2}$$

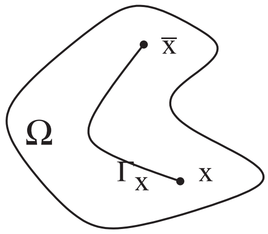


FIGURE 70.2. A curve Γ_x in Ω joining $\bar{x}$ and x.

We first prove that $\varphi(x)$ is independent of the choice of curve Γ_x from $\bar{x}$ to x. Assume that Γ_x and $\tilde{\Gamma}_x$ are two curves from $\bar{x}$ to x. Together they form a closed curve Γ bounding a surface S so Stokes theorem implies

$$\int_\Gamma u \cdot ds = \pm \int_S (\nabla \times u) \cdot n \, ds = 0,$$

since $\nabla \times u = 0$ on S. Now

$$\int_\Gamma u \cdot ds = \int_{\Gamma_x} u \cdot ds - \int_{\tilde{\Gamma}_x} u \cdot ds$$

if we orient Γ in the same direction as Γ_x and thus in the opposite direction as $\tilde{\Gamma}_x$. We conclude that

$$\int_{\tilde{\Gamma}_x} u \cdot ds = \int_{\Gamma_x} u \cdot ds,$$

and the independence of the choice of curve connecting $\bar{x}$ with x follows.

Next, we prove that the function $\varphi(x)$ defined by (70.2) satisfies $\nabla\varphi(x) = u(x)$ for $x \in \Omega$. We do this by choosing a curve Γ_x to connect to x along the x_1-axis, the x_2-axis, or the x_3-axis. Letting Γ_x connect along the x_1-axis according to Fig. 70.3, for $\hat{x}$ close to x we have

$$\varphi(x) - \varphi(\hat{x}) = \int_{\hat{x}_1}^{x_1} u_1(t, x_2, x_3) \, dt \,,$$

and the Fundamental Theorem implies

$$\frac{\partial \varphi}{\partial x_1}(x) = u_1(x).$$

Similarly, we obtain $\frac{\partial \varphi}{\partial x_i}(x) = u_i(x)$ for $i = 2, 3$. We summarize:

Theorem 70.1 *If $u : \Omega \to \mathbb{R}^d$, with Ω being a convex domain in $\mathbb{R}^d$ for $d = 2, 3$, satisfies $\nabla \times u(x) = 0$ for all $x \in \Omega$, then there is a function $\varphi : \Omega \to \mathbb{R}$ such that $u(x) = \nabla\varphi(x)$ for $x \in \Omega$.*

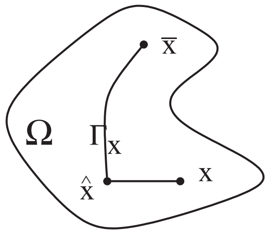


FIGURE 70.3. A curve Γ_x in Ω connecting to x along the x_1-axis.

70.3 A counter-example for a non-convex Ω

Consider the function $u : \Omega \to \mathbb{R}^2$, defined by $u(x) = (-x_2, x_1)/\|x\|^2$ with $\Omega = \{x \in \mathbb{R}^2 : x \neq 0\}$. This function satisfies

$$\nabla \times u(x) = \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2} = \frac{-2x_1x_2}{\|x\|^4} - \frac{-2x_1x_2}{\|x\|^4} = 0 \quad \text{for } x \in \Omega.$$

Nevertheless, $u(x)$ cannot be written in the form $u(x) = \nabla\varphi(x)$ for $x \in \Omega$. This follows by noting that if, for example, Γ is the closed circle given by $s(t) = r(\cos(t), \sin(t))$, $0 \leq t < 2\pi$, then

$$\int_\Gamma u \cdot ds = \int_0^{2\pi} \frac{1}{r^2} r^2 \, dt = 2\pi,$$

while if $u(x) = \nabla\varphi(x)$, $\int_\Gamma u \cdot ds = 0$ since Γ is closed. The reason is that in this case Ω *is not convex*. The point $x = 0$ does not belong to Ω and thus Ω has a "hole". We cannot extend Ω to include $x = 0$ since the function $u(x)$ is singular at $x = 0$ and in particular not Lipschitz continuous at $x = 0$.

Chapter 70 Problems

70.1. If possible, find a potential φ for (a) $u(x) = (x_1, x_2, x_3)$ (b) $u(x) = (x_3, x_1, x_2)$ (c) $u(x) = (x_2^2 - x_3, 2x_1x_2, 3x_3^2 - x_1)$.

70.2. We recall from above that $\nabla \times u = 0$ if and only if $u = \nabla\varphi$ for some φ.

We now ask the question if $\nabla \cdot u = 0$ if and only if $u = \nabla \times \psi$ for some (vector) potential ψ. Recall that we already know that the "if-part" of this is true, namely that $\nabla \cdot u = 0$ if $u = \nabla \times \psi$ for some ψ for some ψ.

It turns out that also the "only if-part" is true, that is, if $\nabla \cdot u = 0$ we may construct a (vector) potential ψ such that $u = \nabla \times |psi$. Verify this, using the construction $\psi(x) = \int_0^1 u(tx) \times tx \, dt$, and assuming $\nabla \cdot u$ in all of $\mathbb{R}^3$ for simplicity.

70.3. Extend the above counterexample to the function $u : \mathbb{R}^3 \to \mathbb{R}^3$ given by $u(x) = (-x_2, x_1, 0)/\|x\|^2$ representing the magnetic field around a current along the x_3-axis.

71
Center of mass and Archimedes' principle*

The simplest schoolboy is now familiar with facts for which Archimedes would have sacrificed his life. (Ernest Renan)

.

71.1 Introduction

We now turn to a study of the stability of floating bodies, including the question of how to design a big ship or a sailing boat so that it does not tip over. An example of an unfortunate design is given by the warship Vasa, which tipped over on its maiden voyage on August 10 1628 in the harbor of Stockholm and sank along with 50 of the crew of 150 people. In the resulting trial, it was decided that the ship was "well built, but badly proportioned" and no-one was held guilty for the disaster. The ship can now be studied at the Vasa museum in Stockholm.

Evidently, the stability properties of Vasa came as a surprise. Vasa had a new design with two gundecks with heavy artillery instead of one and the planned ballast of stone was not sufficient as a counterbalance. The old rules of ship design apparently did not apply to the new design and Calculus and scientific computing at that time was too primitive for trustworthy predictions.

Let's see what we can do today with a little bit of Calculus. We start with the concept of *center of mass*, pass on to *Archimedes principle* and the question of stability of floating bodies.

FIGURE 71.1. Vasa goes unstable.

71.2 Center of mass

Consider a body B occupying the volume V in $\mathbb{R}^3$. Suppose the *density* of the body at x is given by $\rho(x)$. The total mass $m(B)$ of the body is

$$m(B) = \int_V \rho(x)\,dx.$$

The *center of mass* $\bar{x} = (\bar{x}_1, \bar{x}_2, \bar{x}_3) \in \mathbb{R}^3$ of the body B is defined by

$$\bar{x}_i \int_V \rho(x)\,dx = \int_V x_i \rho(x)\,dx\,, \quad i = 1, 2, 3,$$

that is

$$\bar{x}_i = \frac{\int_V x_i \rho(x)\,dx}{\int_V \rho(x)\,dx}\,, \quad i = 1, 2, 3.$$

In vector form, this is

$$\bar{x} = \frac{\int_V x\rho(x)\,dx}{\int_V \rho(x)\,dx}.$$

We now explain the relevance of the concept of center of mass using the concept of *torque*. Assume the body B is acted upon by a vertical gravity force field $-e_3$ of unit strength with the coordinate direction e_3 oriented vertically upward. The torque about a point $\bar{x}$ of a force F acting at x is equal to

$$(x - \bar{x}) \times F = -F \times (x - \bar{x}),$$

see Fig. 71.2. In other words, the torque is a vector that is perpendicular to the plane generated by a force F about $\bar{x}$ and the lever arm $x - \bar{x}$, with modulus equal to the modulus of F times the distance of the point $\bar{x}$ from the line of action of F.

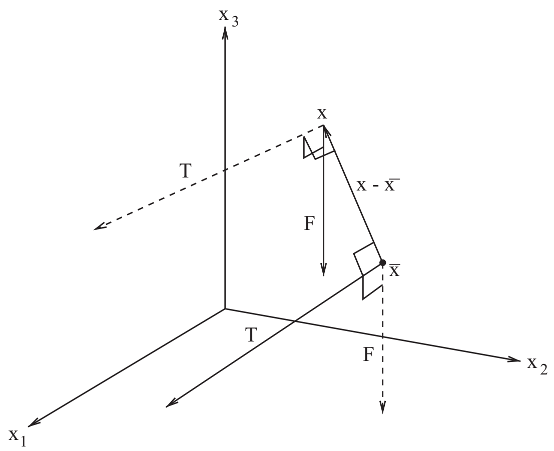


FIGURE 71.2. The torque $T = (x - \bar{x}) \times F$ about the point $\bar{x}$ of F acting at x.

The torque of the gravity field (assuming the acceleration of gravity $g = 1$) acting on an element of mass $\rho(x)\,dx$ at position x about a given point $\bar{x}$ is equal to

$$\rho(x)\,dx\,e_3 \times (x - \bar{x}).$$

The total torque T of the gravity field $-e_3$ on the body B about $\bar{x}$ is thus equal to

$$T = e_3 \times \int_V \rho(x)(x - \bar{x})\,dx = 0,$$

by the definition of the center of mass $\bar{x}$. The torque about $\bar{x}$ thus vanishes which means that body will balance if supported at $\bar{x}$, see Fig. 71.3.

More precisely,

$$T = e_3 \times (\int_V \rho(x)x\,dx - \bar{x}\int_V \rho(x)\,dx) = 0 \tag{71.1}$$

if and only if

$$\bar{x}_i = \frac{\int_V x_i\rho(x)\,dx}{\int_V \rho(x)\,dx},$$

for $i = 1, 2$. This means that the body will balance if supported at a point $x = (x_1, x_2, x_3)$ with $x_1 = \bar{x}_1$ and $x_2 = \bar{x}_2$, while x_3 may be chosen arbitrarily, see Fig. 71.3. Thus, if the body is supported at its center of mass $\bar{x}$ then it will balance independently of its orientation. If the body is supported at a point x different from the center of mass $\bar{x}$, then it will balance only if $\bar{x} - x$ is parallel to the direction of the gravity field.

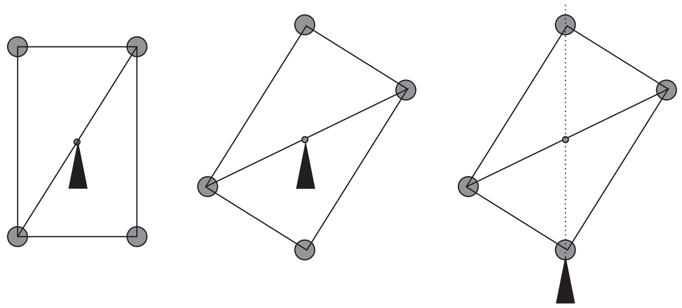


FIGURE 71.3. A body supported at its center of mass, in two stable positions, and a body supported at a boundary point, balanced but unstable.

EXAMPLE 71.1. We compute the center of mass $\bar{x}$ of a thin triangular plate of uniform thickness occupying the region $\Omega = \{x \in \mathbb{R}^2 : 0 \leq x_1, x_2, x_1 + x_2 \leq 1\}$ in the plane. We get

$$\bar{x}_i = \frac{\int_\Omega x_i\, dx}{\int_\Omega dx} = \frac{1/6}{1/2} = \frac{1}{3}.$$

EXAMPLE 71.2. We compute the center of mass of the half-ball $\Omega = \{x \in \mathbb{R}^3 : \|x\| \leq 1,\ x_3 \geq 0\}$. By symmetry $\bar{x}_1 = \bar{x}_2 = 0$. For $\bar{x}_3$ we get using spherical coordinates

$$\begin{aligned}
\int_\Omega x_3\, dx &= \int_0^{2\pi} \int_0^{\pi/2} \int_0^1 r\cos(\varphi)\, r^2 \sin(\varphi)\, dr d\varphi d\theta \\
&= \int_0^{2\pi} \int_0^{\pi/2} \frac{1}{2}\sin(2\varphi) \left[\frac{1}{4} r^4\right]_0^1 d\varphi d\theta \\
&= \frac{1}{4}\int_0^{2\pi} \left[\frac{1}{4}\cos(2\varphi)\right]_0^{\pi/2} d\theta = \frac{1}{4}\int_0^{2\pi} d\theta = \pi/4,
\end{aligned}$$

that is, $\bar{x}_3 = \int_\Omega x_3\, dx / \int_\Omega dx = \frac{\pi/4}{2\pi/3} = \frac{3}{4}$.

71.3 Archimedes' principle

Archimedes principle states that (i) the *buoyancy force* acting on a body B totally immersed in a liquid is equal to the weight of the displaced liquid and (ii) acts along a vertical line through the center of mass of the displaced fluid, which we refer to as the *center of bouyancy* c_b. We shall now prove this fact using vector Calculus. The force from the fluid acting on an element $ds = ds(x)$ of the surface S of the body B at position x is equal to $-p(x)n(x)\,ds$, where $p(x)$ is the pressure of the liquid and $n(x)$ is the outward (from B) unit normal to S at x. The total pressure force on B is thus

$$F = -\int_S p(x)n(x)\,ds(x) = -\int_S pn\,ds.$$

Since

$$\int_V \frac{\partial p}{\partial x_i}\,dx = \int_S pn_i\,ds, \quad i = 1, 2, 3,$$

where V is the volume occupied by B, we have

$$F = -\int_V \nabla p(x)\,dx$$

The *pressure* $p(x)$ in a fluid at rest, called the *hydro-static pressure*, is given by

$$p(x) = \rho z(x) + p_0,$$

where $z(x)$ is the depth, ρ is the constant density of the fluid and p_0 is the pressure on the surface of the fluid, see Fig. 71.4.

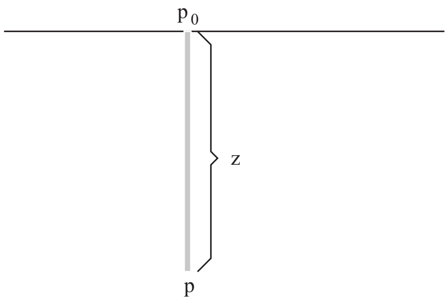


FIGURE 71.4. Hydrostatic pressure $p(x) = \rho z(x) + p_0$.

The pressure force at a point x is equal in all directions and its modulus $p(x)$ equal is to the weight $\rho z(x)$ of the column of fluid above the point x plus the surface pressure p_0 from the atmosphere. We conclude that

$$\nabla p(x) = -\rho e_3,$$

where we assume that the coordinate direction e_3 is vertical and pointed upwards. Therefore,

$$F = \int_V \rho \, dx \, e_3 \equiv W e_3,$$

where $W = \int_V \rho \, dx$ is the total weight of the displaced fluid. This proves the first part of Archimedes principle.

Next, the total torque T from the fluid pressure forces on S about a point $\bar{x}$ is given by

$$T = \int_S (x - \bar{x}) \times (-p(x)n(x)) \, ds(x) = \int_S n(x) \times p(x)(x - \bar{x}) \, ds.$$

Recalling that

$$\int_S n \times F \, ds = \int_V \nabla \times F \, dx,$$

we find that

$$T = \int_V \nabla \times (p(x)(x - \bar{x})) \, dx.$$

But,

$$\nabla \times (p(x)(x - \bar{x})) = \nabla p \times (x - \bar{x}) + p \, \nabla \times (x - \bar{x}).$$

Since $\nabla \times (x - \bar{x}) = 0$, it follows that

$$T = \int_V \nabla p \times (x - \bar{x}) \, dx = -\int_V \rho(x - \bar{x}) \, dx \times e_3$$

and the torque T vanishes if $\bar{x}$ satisfies

$$\bar{x}_i \int_V \rho \, dx = \int_V x_i \rho \, dx \quad \text{for } i = 1, 2.$$

We conclude that the the buoyancy force is vertical upward and is acting along a vertical line through the center of mass of the displaced fluid. We have now proved:

Theorem 71.1 (Archimedes' principle) *The buoyancy force acting on a body immersed in a liquid is equal to the weight of the displaced liquid and acts along a vertical line through the center of mass of the displaced fluid.*

We can directly extend Archimedes' principle to a partially immersed body assuming the pressure on the surface of the fluid to be is zero.

71.4 Stability of floating bodies

The stability of a floating body B is of central importance in all forms of boating, from canoes to big ships. The question of stability can be reduced to a question of the relative position of (i) the center of mass c_m of the body B and (ii) the center of buoyancy c_b of B according to the following discussion. Consider the body in rest position with the gravity force acting vertically downward from the center of gravity, and the buoyancy force acting vertically upward from the center of buoyancy. We assume the body is in equilibrium with the gravity force and the buoyancy force balancing and acting along the vertical line through the centers of gravity and buoyancy, see Fig. 71.5.

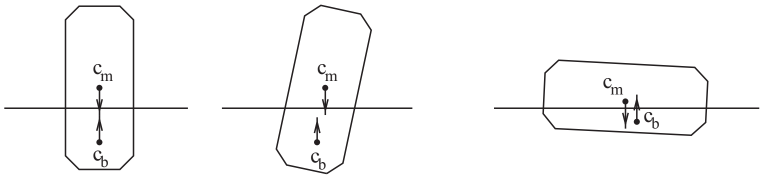


FIGURE 71.5. Floating bodies with centers of gravity and buoyancy.

Assume that the body is tilted a small angle so that the centers of gravity and buoyancy are displaced horizontally, see Fig. 71.5. Let T be the resulting torque from the pair of gravity and buoyancy forces. The sign of T will govern the stability! If T acts in the same direction as the tilting, then the tendency of tilting will be enforced and the body will depart from its equilibrium position and eventually tilt over, see Fig. 71.5. This happens if the center of gravity is displaced horizontally in the direction of tilting more quickly than the center of buoyancy. Conversely, if the center of gravity is displaced more slowly, then the resulting torque T will be negative and act as a restoring force seeking to bring back the body to the rest position, see Fig. 71.5. We now consider two examples with simple geometry.

Example 71.3. Consider a space capsule with hemispherical base and conical top floating in the Pacific and waiting to be recovered. Will the capsule float upright or not? Assuming the capsule is floating upright with a part of the hemispherical base immersed into the water, see Fig. 71.6.

The resultant of the buoyancy forces is directed upward and acts through the center C of the hemisphere, see Fig. 71.6. If the capsule is tilted a little, the resultant of the buoyancy forces is still directed upwards through C and the torque from the gravity force will be de-stabilizing if the center of mass c_m of the capsule is positioned above C, and sta-

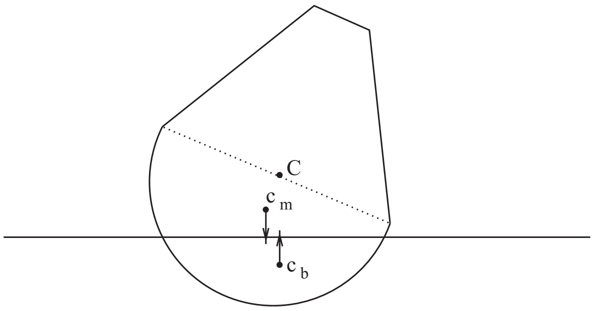


FIGURE 71.6. Space capsule floating upright.

bilizing if c_m is below C, for c_m on the symmetry axis of the capsule, see Fig. 71.6.

EXAMPLE 71.4. Consider a rectangular box with square horizontal cross section of width $2w$ and height $2h$ and density $\bar{\rho}$ which is floating in a fluid of density ρ, see Fig. 71.7. Suppose that $\bar{\rho}$ is small compared to ρ so that it penetrates into the fluid only slightly. To test the stability of the box, suppose the box is rotated a small angle θ around the mid-point C at the bottom. The de-stabilizing torque about C resulting from the gravity force through the center of gravity is equal to $g\bar{\rho}(2w)^2 2hh\sin(\theta)$, see Fig. 71.7. The stabilizing torque from the change of buoyancy forces caused by the rotation is equal to

$$2\frac{2}{3}www\sin\theta\frac{1}{2}\rho g w,$$

because the area of the triangle CAB is equal to $ww\sin\theta\frac{1}{2}$, and the center of gravity of CAB is at horizontal distance $2\frac{2}{3}w$ from C. The position is stable if

$$2\frac{2}{3}w^4\sin\theta\frac{1}{2}\rho g > g\bar{\rho}8w^2h^2\sin(\theta),$$

that is, if

$$w^2\rho > 12h^2\bar{\rho}.$$

Chapter 71 Problems

71.1. The density of ice is 0.917 times the density of water (at $-4°$C). How large a part of an iceberg is visible above the water surface?

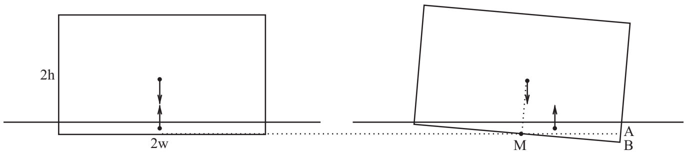


FIGURE 71.7. Floating box.

71.2. How does a log float? Why does it not want to float in an upright position?

71.3. Understand why catamaran ships have good stability properties.

71.4. Find the stable floating position of a "log" with a quadratic cross-section and density $\bar{\rho} = \frac{1}{2}\rho$. Find the stable positions as a function of the ratio $\bar{\rho}/\rho$ (we know from above that for $\bar{\rho}/\rho$ sufficiently small it will float as the box in Fig. 71.7). May there be more than one stable position (disregarding symmetric ones). Discuss! May the conclusion depend on the shape of the cross-section?

71.5. How does a (perfect) ice cube float? How does a barrel (cylinder) float, given hight/diameter ratio and density?

71.6. Study the design of sailing boats from the stability point of view. Study in particular modern designs with good form stability (wide and flat bottom), and classical designs with a narrow deep hull. Connect to the discussion above.

71.7. Extend Archimedes principle to a body immersed into a system of two layers of different fluids on top of each other.

72
Newton's nightmare*

> God does not care about mathematical difficulties. He integrates empirically. (Einstein)

Newton's theory of gravitation states that the gravitational force field $F(x)$ generated by a point mass m at the origin is the potential field

$$F(x) = -m\frac{x}{\|x\|^3} = \nabla(\frac{m}{\|x\|}), \tag{72.1}$$

corresponding to the potential $\varphi(x) = m/\|x\|$, in units where the gravitational constant is one. This means the gravitational force from the mass m at the origin on a unit point mass at position x is equal to $F(x)$. Taking norms gives

$$\|F(x)\| = \frac{m}{\|x\|^2},$$

which is known as *Newton's Inverse Square Law* . More generally, the gravitational force field of a mass m a position y is given by

$$F(x) = -m\frac{x-y}{\|x-y\|^3}, \tag{72.2}$$

with $F(x)$ being the force on a unit point mass a position x and the corresponding potential $\varphi(x) = m/\|x-y\|$.

Over a long period, Newton tried to show one consequence of his new theory of gravitation: the gravitational force between two solid balls is the same as if the total mass of each ball was concentrated at the center of

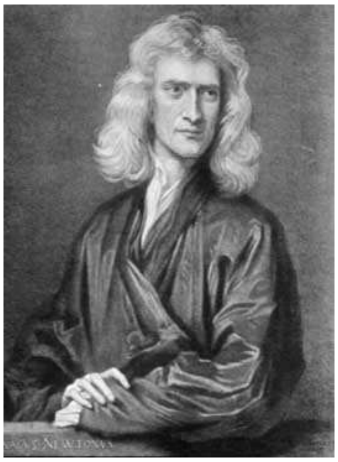


FIGURE 72.1. Isaac Newton 1689: "I have not been able to discover the cause of those properties of gravity from phenomena, and I frame no hypotheses; for whatever is not deduced from the phenomena is to be called a hypothesis, and hypotheses, whether metaphysical or physical, whether of occult qualities or mechanical, have no place in experimental philosophy".

mass of each ball. This result has important practical implications. For example, it would allow the modeling of the solar system as 9 small point masses representing the planets orbiting around one fixed big point mass representing the Sun, that is as a 9-body system. Without the simplifying basic result, we would have to take into account the gravitational attraction between the parts of each of the bodies and we would end up with a very complicated model. The practicality of Newton's gravitational theory could easily be questioned by anyone having some interest in that direction, like the Church. Lacking this basic result, Newton delayed the publication of his monumental *Principia Mathematica* many years. Newton states that he purposely made Principia difficult to read "to avoid being bated by little smatterers of mathematics". Newton did not like critics.

It fact, even a 9-body system of point masses may be far beyond comprehension or mathematical analysis. Luckily, the solar system is a very special 9-body system in which the motion of each planet can be viewed to good approximation as a 1-body system, i.e. as each planet orbiting undisturbed around one heavy Sun. Such 1-body systems have a full analytical solution available, as we saw in the Chapter Lagrange and the Principle of Least Action.

The basic result that Newton finally succeeded in proving can be phrased as follows: Consider a thin spherical shell S of radius r and uniform thickness centered at the origin and assume the total mass of the shell is m. Let $F(x)$ be the gravitational force field generated by the spherical shell so that $F(x)$ is the gravitational force of the shell on a unit point mass at position x outside the sphere. Newton proved that

$$F(x) = -m\frac{x}{\|x\|^3} \quad \text{for } \|x\| > r,$$

which says that the gravitational field generated by the sphere on a point outside the sphere is the same as the field generated by a point mass m at the center of the sphere.

The gravitational field $F(x)$ of the shell/sphere is the sum of the gravitational fields of all the little pieces $ds(y)$ of the surface of mass $dm(y)$ at position y making up the sphere S, that is

$$F(x) = \int_S f(y)ds(y),$$

where

$$f(y)ds(y) = -dm(y)\frac{x-y}{\|x-y\|^3}$$

is the gravitational field of the piece of surface $ds(y)$ of mass $dm(y)$ at position y. We note that

$$dm(y) = \frac{mds(y)}{4\pi r^2},$$

since the area of the sphere is $4\pi r^2$ and the total mass is m, and thus

$$f(y) = -\frac{m}{4\pi r^2}\frac{x-y}{\|x-y\|^3}. \tag{72.3}$$

Newton thus wanted to verify that

$$\int_S f(y)\,ds(y) = -m\frac{x}{\|x\|^3} \quad \text{for } \|x\| > r, \tag{72.4}$$

where $f(y)$ is given by (72.3). Once this basic result for a sphere is established, the corresponding result for a solid ball follows by simply viewing the ball as the union of a collection of thin spheres of varying radii. The desired final result for two solid balls follows similarly.

We now prove (72.4) giving the gravitational field of a thin spherical shell S of radius r and total mass m centered at the origin. We assume that $x = (R, 0, 0)$ with $R > r$. By symmetry, this covers the general situation. We note that the components $F_2(x)$ and $F_3(x)$ of the gravitational force vanish because the gravitational force is directed from $(R, 0, 0)$ towards the origin, and we have simply to verify that

$$F_1(x) = -\frac{m}{4\pi r^2}\int_S \frac{R-y_1}{\|x-y\|^3}ds(y) = -\frac{m}{R^2}.$$

To compute the surface integral, we use spherical coordinates

$$y = (r\cos(\varphi), r\sin(\varphi)\cos(\theta), r\sin(\varphi)\sin(\theta))$$

with $0 \le \varphi \le \pi$ and $0 \le \theta \le 2\pi$, see Fig. 72.2, and recall from the Chapter Surface integrals that

$$ds(y) = r^2 \sin(\varphi) d\varphi\, d\theta.$$

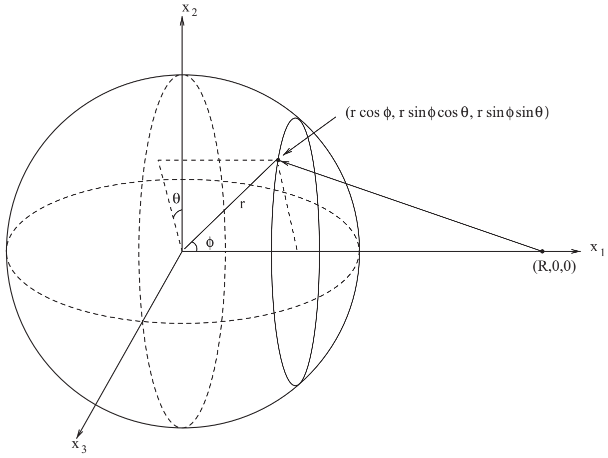


FIGURE 72.2. Newtons nightmare:

We have according to Fig. 72.2,

$$\begin{aligned}
F_1(x) &= -\frac{m}{4\pi r^2}\int_S \frac{R-y_1}{\|x-y\|^3} ds(y) \\
&= -\frac{m}{4\pi}\int_0^\pi \int_0^{2\pi} \frac{(R-r\cos(\varphi))\sin(\varphi)}{((R-r\cos(\varphi))^2+(r\sin(\varphi))^2)^{3/2}}\, d\theta d\varphi \\
&= -\frac{m}{2}\int_0^\pi \frac{(R-r\cos(\varphi))\sin(\varphi)}{((R-r\cos(\varphi))^2+(r\sin(\varphi))^2)^{3/2}}\, d\varphi
\end{aligned}$$

where we performed the integration with respect to θ using the fact that the integrand is independent of θ. We thus need to verify that

$$I = \int_0^\pi \frac{(R-r\cos(\varphi))\sin(\varphi)\, d\varphi}{((R-r\cos(\varphi))^2+(r\sin(\varphi))^2)^{3/2}} = \frac{2}{R^2}. \tag{72.5}$$

To this end, we change variables to set $t = \cos(\varphi)$ and we use $dt = -\sin(\varphi)d\varphi$ to get

$$I = \int_{-1}^1 \frac{(R-rt)\, dt}{(R^2+r^2-2Rrt)^{3/2}} = \frac{1}{R^2}\int_{-1}^1 \frac{(1-at)\, dt}{(1+a^2-2at)^{3/2}},$$

where $a = \frac{r}{R} < 1$. By a routine computation, we find that $a < 1$ implies

$$\begin{aligned}
&\int_{-1}^{1} \frac{(1-at)\,dt}{(1+a^2-2at)^{3/2}} \\
&= \int_{-1}^{1} \frac{(\frac{1+a^2}{2}-at)\,dt}{(1+a^2-2at)^{3/2}} - \int_{-1}^{1} \frac{(\frac{1+a^2}{2}-1)\,dt}{(1+a^2-2at)^{3/2}} \\
&= \frac{1}{2a}\big[-(1+a^2-2at)^{1/2}\big]_{-1}^{1} - \frac{a^2-1}{2a}\big[(1+a^2-2at)^{-1/2}\big]_{-1}^{1} \\
&= \frac{1}{2a}\big(1+a-(1-a)\big) - \frac{a^2-1}{2a}\big(\frac{1}{1-a}-\frac{1}{1+a}\big) = 1+1 = 2,
\end{aligned}$$

and the desired result follows:

$$F_1(x) = -\frac{m}{R^2} \quad \text{if } x = (0,0,R), \quad R > r.$$

Below, we give a much shorter proof of this result using some tools of Calculus to be developed in the next chapters.

Chapter 72 Problems

72.1. Prove that the gravitational field from a thin sphere is equal to zero *inside* the sphere.

72.2. Compute the gravitational field $F(x)$ for $x \in \mathbb{R}^3$ of a solid ball of total mass m and radius r centered at the origin

72.3. Compute the gravitational field of a "black hole" with mass density $\frac{exp(-r)}{r}$, $r = \|x\|$.

72.4. Determine the gravitational field generated by a thin straight uniform rod.

72.5. Determine the gravitational field generated by a thin circular flat (a) ring (b) disc.

72.6. (a) Consider a particle cloud of uniform density in the form of a ball. Assume the particles attract each other according to Newton's Law of gravitation. Compute the evolution of the cloud for $t > 0$ assuming the particles are at rest at $t = 0$. (b) Do the same with a cloud in the form of the volume between two concentric spheres. (c) Extend to clouds of variable density.

73
Laplacian models

> ... on aura donc $\Delta u = 0$; cette équation remarquable nous sera de la plus grande utilité .... (Laplace)
>
> If one has to stick to this damned quantum jumping, then I regret ever having been involved in this thing. I don't like it (quantum mechanics), and I'm sorry I ever had anything to do with it. (Schrödinger)

73.1 Introduction

In this chapter, we present some basic models involving the Laplacian, including models for heat conduction, elasticity, electromagnetics, fluid mechanics, and gravitation. In deriving these models, we make use of the basics of Calculus in several dimensions including Gauss' and Stokes' theorems, and we get a quick and easy introduction to some of mysteries of the mechanics and physics of "continuous media". We also make connections to linear algebra when discretizing the Laplacian using the 5-point scheme and variants of "Svensson's formula".

73.2 Heat conduction

We first model *heat conduction* in a heat-conducting material occupying the volume Ω in $\mathbb{R}^3$ with boundary Γ, over a time interval $I = [0, T]$. We let $u(x,t)$ denote the *temperature* and $q(x,t)$ the *heat flux* at the point x at time t. The heat flux is a vector $q = (q_1, q_2, q_3)$, where q_i is the heat flux, or rate of heat flowing in the direction x_i. We let $f(x,t)$ denote the rate of heat (per unit of volume) supplied at (x,t) by a *heat source*.

We derive the model using a basic *conservation law* expressing *conservation of heat* in the following form: for any fixed domain V in Ω with boundary S, the rate of the total heat introduced in V by the external source is equal to the rate of the total heat accumulated in V plus the total heat flux through S. This is based on the conviction that the heat introduced in V by the external source can choose between two options only: (i) flow out of V or (ii) be accumulated in V. With S denoting the boundary of V and n denoting the outward unit normal to S, see Fig. 73.1, the conservation law can be expressed as

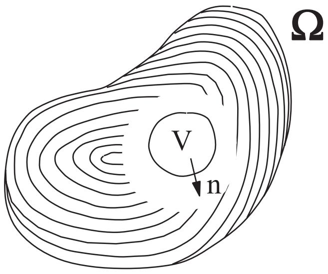


FIGURE 73.1. An arbitrary subset V of a heat conducting body Ω.

$$\int_V f\,dx = \frac{\partial}{\partial t}\int_V \lambda u\,dx + \int_S q\cdot n\,ds, \tag{73.1}$$

where $\lambda(x,t)$ is the *heat capacity coefficient* giving the amount heat per unit of volume needed to raise the temperature one unit, and all functions are evaluated at a specific time $t \in I$. By the Divergence theorem,

$$\int_S q\cdot n\,ds = \int_V \nabla\cdot q\,dx,$$

and combined with (73.1), this implies that

$$\int_V \left(\frac{\partial}{\partial t}(\lambda u) + \nabla\cdot q\right)\,dx = \int_V f\,dx,$$

where the time derivative could be moved under the integral sign because V does not depend on time t. Since V is arbitrary, assuming the integrands are Lipschitz continuous, it follows that

$$\frac{\partial}{\partial t}(\lambda u)(x,t) + \nabla\cdot q(x,t) = f(x,t) \quad \text{for all } x\in\Omega,\ 0<t\le T, \tag{73.2}$$

which is a differential equation describing *conservation of heat* involving two unknowns: the temperature $u(x,t)$ and the heat flux $q(x,t)$. We thus have one equation and two unknowns and we need yet another equation.

The second equation is a *constitutive equation* that couples the heat flux q to the temperature gradient ∇u. *Fourier's law* states that heat flows from warm to cold regions with the heat flux proportional to the temperature gradient:

$$q(x,t) = -a(x,t)\nabla u(x,t) \quad \text{for } x \in \Omega,\ 0 < t \leq T \tag{73.3}$$

where the factor of proportionality $a(x,t)$ is the coefficient of heat conductivity. Note the minus sign indicating that the heat flows from warm to cold regions, and that the heat conductivity $a(x,t)$ is positive. Combining (73.2) and (73.3), we obtain the basic differential equation describing heat conduction:

$$\frac{\partial}{\partial t}(\lambda u) - \nabla \cdot (a\nabla u) = f \quad \text{in } \Omega \times (0,T], \tag{73.4}$$

where $a(x,t)$ and $\lambda(x,t)$ are given positive coefficients depending on (x,t) and $f(x,t)$ is a given heat source, and the unknown $u(x,t)$ represents the temperature.

To define the solution uniquely, the differential equation is complemented by initial and boundary conditions. The complete model with *Dirichlet boundary conditions* reads

$$\begin{cases} \frac{\partial}{\partial t}(\lambda u) - \nabla \cdot (a\nabla u) = f & \text{in } \Omega \times (0,T], \\ u = u_b & \text{on } \Gamma \times (0,T], \\ u(x,0) = u_0(x) & \text{for } x \in \Omega, \end{cases} \tag{73.5}$$

where u_0 is the initial temperature and u_b is the boundary temperature. The Dirichlet boundary condition corresponds to immersing the body Ω in a large reservoir with a specified temperature u_b and assuming that the boundary acts as a perfect thermal conductor so that the temperature of the body on the boundary is equal to the specified outside reservoir temperature u_b. Note that the given boundary temperature $u_b = u_b(x,t)$ may vary with (x,t).

Other commonly encountered boundary conditions are *Neumann* and *Robin* boundary conditions. A Neumann boundary condition corresponds to prescribing the heat flux $q \cdot n$ across (out of) the boundary:

$$q \cdot n = -a\nabla u \cdot n = -a\frac{\partial u}{\partial n} = -a\partial_n u = g \quad \text{on } \Gamma,$$

with g given. A *homogeneous Neumann boundary condition* with $g = 0$ corresponds to a perfectly insulating boundary with the heat flux across the boundary being zero. A homogenous Robin boundary condition is intermediate with the boundary neither being perfectly conducting nor perfectly insulated, with the heat flux through the boundary being proportional to the difference of the temperature u inside and a given temperature u_b outside Ω:

$$-a\partial_n u = \kappa(u - u_b)$$

with κ a positive coefficient representing the heat conductivity of the boundary.

Partitioning the boundary Γ into disjoint pieces Γ_1, Γ_2 and Γ_3 with different types of boundary conditions, the *general initial boundary value problem* IBVP for the heat equation has the form,

$$\begin{cases} \frac{\partial}{\partial t}(\lambda u) - \nabla \cdot (a\nabla u) = f & \text{in } \Omega \times (0,T], \\ u = u_b & \text{on } \Gamma_1 \times (0,T], \\ -a\partial_n u = g & \text{on } \Gamma_2 \times (0,T], \\ a\partial_n u + \kappa(u - u_b) = 0 & \text{on } \Gamma_3 \times (0,T], \\ u(x,0) = u_0(x) & \text{for } x \in \Omega, \end{cases} \tag{73.6}$$

where u_b represents a given "exterior" boundary temperature, and g represents a given outward normal heat flux on the boundary.

We note that in a stationary situation with $\frac{\partial}{\partial t}(\lambda u) = 0$ and with the heat source $f = 0$, the equation (73.2) expressing conservation of heat, takes the form

$$\nabla \cdot q = 0. \tag{73.7}$$

If heat is neither produced nor accumulated, then conservation of heat is expressed by the equation $\nabla \cdot q = 0$, that is, the heat flux q is *divergence-free*. Below we shall meet several other examples of divergence-free fields.

73.3 The heat equation

We refer to the special case of (73.6) with $\lambda = a = 1$ as the *heat equation*. In the case with homogeneous Dirichlet boundary conditions, we get the model

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = f & \text{in } \Omega \times (0,T], \\ u = 0 & \text{on } \Gamma \times (0,T], \\ u(x,0) = u_0(x) & \text{for } x \in \Omega, \end{cases}$$

where u_0 is the initial temperature, and $\Delta u = \nabla \cdot (\nabla u)$ is the Laplacian. The heat equation serves as a basic prototype of a *parabolic problem*.

FIGURE 73.2. Poisson

73.4 Stationary heat conduction: Poisson's equation

The stationary analog of (73.6) reads

$$\begin{cases} -\nabla \cdot (a\nabla u) = f & \text{in } \Omega, \\ u = u_b & \text{on } \Gamma_1, \\ -a\partial_n u = g & \text{on } \Gamma_2, \\ a\partial_n u + \kappa(u - u_b) = 0 & \text{on } \Gamma_3. \end{cases} \tag{73.8}$$

Choosing $a = 1$ leads to the *Poisson equation*

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = u_1 & \text{on } \Gamma_1, \\ -a\partial_n u = g_2 & \text{on } \Gamma_2, \\ a\partial_n u + \kappa(u - u_b) = g_3 & \text{on } \Gamma_3. \end{cases} \tag{73.9}$$

In the case of homogeneous Dirichlet boundary conditions on the whole of the boundary, the Poisson equation reads

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma. \end{cases} \tag{73.10}$$

Poisson's equation serves as a basic model of an *elliptic* problem and has numerous applications in physics and mechanics. We present the basic applications below. Poisson's equation $-\Delta u = f$ with $f = 0$ is referred to as *Laplace's equation*: $\Delta u = 0$.

We now give a couple of analytic solutions to the heat equation in simple situations:

EXAMPLE 73.1. The stationary temperature u in a heat conduction unit cube Q with heat production and conduction coefficient equal to one, zero boundary temperature for $x_1 = 0, 1$, and zero heat flux for $x_2, x_3 = 0, 1$, is given by

$$u(x) = \frac{1}{2}(x_1(1 - x_1).$$

We see that the temperature is maximal for $x_1 = 0.5$ and drops off quadratically towards the Dirichlet boundary, see Fig. 73.3 for a plot in the corresponding case in two dimension in the unit square.

EXAMPLE 73.2. Consider the homogenous heat equation in the unit square Q with $f = 0$ and homogenous Dirichlet boundary conditions: the function

$$u(x, t) = e^{-(n^2+m^2)t} \sin(nx_1) \sin(mx_2)$$

with $m, n = 1, 2, 3, \ldots$ is a solution of the homogenous heat equation $\frac{\partial u}{\partial t} - \Delta u = 0$ with initial value $u_0(x_1, x_2) = \sin(nx_1) \sin(mx_2)$, see Fig. 73.3. We see that the temperature $u(x, t)$ decays exponentially in time very quickly if n and/or m is only moderately large. This corresponds to the fact that a temperature oscillating in space, is quickly levelled out.

EXAMPLE 73.3. The stationary temperature $u(x)$ between the two planes $\{x_3 = 0\}$ and $\{x_3 = 1\}$ bounding a heat conducting layer with heat conductivity coefficient equal to one, zero heat source, and the temperature $u = 1$ on $\{x_3 = 1\}$ and $u = 0$ on $\{x_3 = 0\}$, is given by $u(x) = x_3$ displaying a linear variation of the temperature between the plates. No surprise of course.

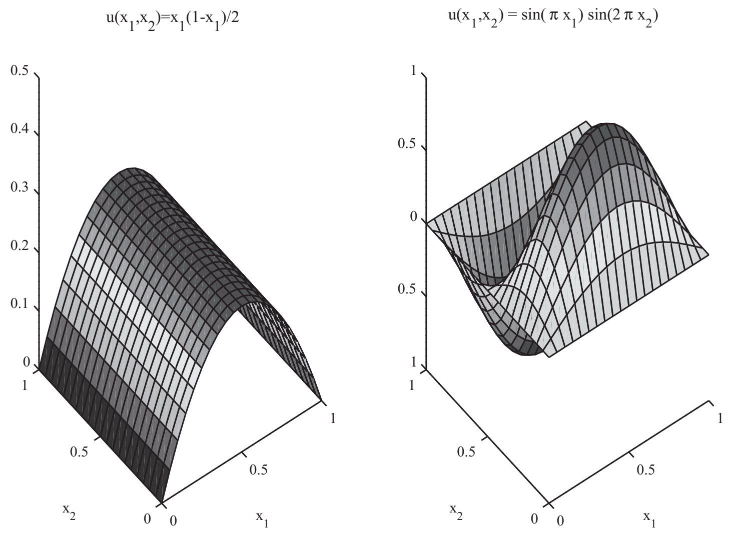


FIGURE 73.3. The functions $\frac{1}{2}(x_1(1-x_1)$ and $sin(\pi x_1)sin(2\pi x_2)$

73.5 Convection-diffusion-reaction

The heat equation models the physical phenomenon of *diffusion*, and we now extend this model to include the phenomena of *convection* and *reaction*. We obtain a scalar *convection-diffusion-reaction* equation, which is another basic model in science. We consider a typical case where u represents the concentration of a certain chemical species subject to convection in a given velocity field $\beta(x,t)$, diffusion with diffusion coefficient $\epsilon(x,t)$ and reaction with reaction rate $\alpha(x,t)$. For example, u may represent the concentration of a contaminant in a volume of water moving with the velocity $\beta(x,t)$.

The model results from a principle of conservation of mass together with a constitutive equation generalizing Fourier's law expressing the *flow rate* q of the chemical species in terms of ∇u and βu. Conservation of mass is expressed by

$$\dot{u} + \nabla \cdot q + \alpha u = f,$$

where f represents a source term, and the constitutive law takes the form

$$q = \beta u - \epsilon \nabla u.$$

which says that the total flow rate q is the sum of a convective rate βu and a diffusive rate $-\epsilon \nabla u$. The model thus takes the form:

$$\dot{u} + \nabla \cdot (\beta u) + \alpha u - \nabla \cdot (\epsilon \nabla u) = f \quad \text{in } \Omega \times (0, T], \tag{73.11}$$

together with initial and boundary conditions, where Ω is domain in space and $[0,T]$ a given time interval. We shall meet this model and generalizations thereof in several different contexts below.

73.6 Elastic membrane

Consider a horizontal elastic net covering the unit square $Q = \{x \in \mathbb{R}^2 : 0 \leq x_i \leq 1, i = 1,2\}$ formed by elastic strings tied together at nodes $a_{ij} \in \mathbb{R}^2$ in a uniform quadrilateral mesh with mesh size $h = 1/N$, so that $a_{ij} = (ih, jh)$, $i, j = 0, 1, ..., N$, where N is the number of cells in each coordinate direction. Assume that the net is stretched so that the tension in each string is equal to h, corresponding to the tension being equal to one per unit of length. Note the normalization introduced says that the tension in each string decreases as the number of strings increases. We refer to the situation in which all the nodes lie in the plane of the square and there is no external load on the net as the unloaded reference configuration of the net.

Suppose the net is subject to a set of downward vertical loads of size $f_{ij}h^2$ at the nodes a_{ij}. The net will deform under the loads and the nodes will be displaced from the initial unloaded reference configuration. Let the vertical displacement of node a_{ij} be denoted by $u_{i,j}$. If the displacements are small, then (recalling the Chapter String theory) the vertical upward force from the net on node a_{ij} is equal to

$$(u_{i,j} - u_{i-1,j}) + (u_{i,j} - u_{i+1,j}) + (u_{i,j} - u_{i,j-1}) + (u_{i,j} - u_{i,j+1}),$$

with contributions from the four pieces of string meeting at a_{ij}. This is because the vertical slope of the line between for example node (i,j) and $(i-1,j)$ is equal to $(u_{i,j} - u_{i-1,j})/h$ and the tension is h. We thus obtain the following vertical equilibrium equation for each node a_{ij}:

$$-\frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2} - \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{h^2} = f_{ij}.$$

Passing to the limit as h tends to zero, and recalling that Taylor's theorem implies

$$\lim_{h \to 0} \frac{v(x-h) - 2v(x) + v(x+h)}{h^2} = v''(x) = \frac{d^2 v}{dx^2}(x)$$

if $v : \mathbb{R} \to \mathbb{R}$ is twice differentiable, we are led to the equation

$$-\Delta u(x) = f(x).$$

This equation expresses the equilibrium of a horizontal membrane made by an elastic fabric and carrying a vertical load of intensity (force per unit area) $f(x)$, where $u(x)$ is the vertical displacement of the membrane at x and we assume that the membrane in its unloaded plane reference configuration is prestressed to uniform tension in all directions.

We can generalize to a horizontal membrane covering a general domain Ω in $\mathbb{R}^2$. Assuming the membrane is fixed at the boundary Γ of Ω, so that the vertical displacement $u(x)$ is zero at Γ, we thus obtain Poisson's equation

$$-\Delta u = f \quad \text{in } \Omega, \; u = 0 \quad \text{on } \Gamma \tag{73.12}$$

as a model for the vertical deflection of a horizontal elastic membrane spanned over the boundary Γ of a domain Ω in $\mathbb{R}^2$, subject to a vertical load of intensity $f(x)$. This is a basic model of elasticity theory.

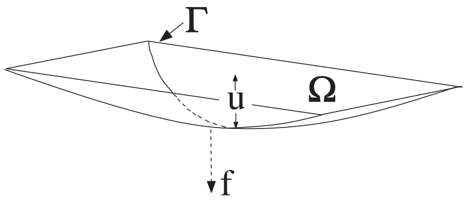


FIGURE 73.4. An elastic membrane under a load $f(x)$ and supported at Γ.

EXAMPLE 73.4. If $\Omega = \{x \in \mathbb{R}^2 : \|x\| < 1\}$ is the unit disc, and the load f is radially symmetric, then the deflection u will also be radially symmetric. Recalling the form of the Laplacian in polar coordinates from the Chapter The divergence, rotation and Laplacian, we can write (73.12) in the form

$$-\Delta u = -\frac{1}{r}\frac{\partial}{\partial r}(r\frac{\partial u}{\partial r}) = f(r) \quad \text{for } 0 < r < 1,\ u(1) = 0, \frac{\partial u}{\partial r}(0) = 0.$$

Note the boundary condition $\frac{\partial u}{\partial r}(0) = 0$, which has no counterpart in x-coordinates, says that $u(x)$ is differentiable at $x = 0$. If $\frac{\partial u}{\partial r}(0) \neq 0$, then $u(x)$ has a a conical "to" at $x = 0$ and thus is not differentiable at $x = 0$. If $f(r) = 1$, then the solution is given by

$$u(r) = \frac{1}{4}(1 - r^2) \quad \text{for } 0 \leq r \leq 1.$$

73.7 Solving the Poisson equation

Suppose we would like to numerically solve the Poisson equation

$$-\Delta u = f \quad \text{in } Q, \qquad u = u_b \quad \text{on } \Gamma$$

where Q is the unit square with boundary Γ and $f(x)$ a given function on Q. Recalling the derivation of the model $-\Delta u = f$ from the previous section, we are led to computing approximations $U_{i,j}$ of $u(ih, jh)$ for $i, j = 0, 1, ..., N$, where $h = 1/N$, from the system of equations

$$-\frac{U_{i-1,j} - 2U_{i,j} + U_{i+1,j}}{h^2} - \frac{U_{i,j-1} - 2U_{i,j} + U_{i,j+1}}{h^2} = f(ih, jh),$$
$$i, j = 1, ..., N-1,$$

that is

$$4U_{i,j} - U_{i-1,j} - U_{i+1,j} - U_{i,j-1} - U_{i,j+1} = h^2 f(ih, jh),$$
$$i, j = 1, ..., N-1, \quad (73.13)$$

where $U_{i,j} = u_b(ih, jh)$ if i or j is equal to 0 or N. We see that this is an $m \times m$ system of equations with $m = (N-1) \times (N-1)$ in the unknowns $U_{i,j}$ where $i, j = 1, ..., N-1$. This is the famous *5-point scheme* for the Poisson's equation, where the unknown $U_{i,j}$ is coupled to its four neighbors $U_{i-1,j}$, $U_{i+1,j}$, $U_{i,j-1}$, $U_{i,j+1}$.

If $f = 0$, then the 5-point scheme takes the form ("Svensson's formula")

$$U_{i,j} = \frac{1}{4}(U_{i-1,j} + U_{i+1,j} + U_{i,j-1} + U_{i,j+1}),$$

stating that each value $U_{i,j}$ is the mean value of the neighboring values.

Note that (73.13) is a linear system of equations for the values of U that requires some work to solve. For example, we may try to solve (73.13) by fixed point iteration as follows with $k = 0, 1, ...$

$$U_{i,j}^{k+1} = U_{i,j}^{k}$$
$$- \alpha\big(4U_{i,j}^k - U_{i-1,j}^k - U_{i+1,j}^k - U_{i,j-1}^k - U_{i,j+1}^k - h^2 f(ih, jh)\big), \quad (73.14)$$

for $i, j = 1, ..., N-1$, with $U_{i,j}^{k+1} = u_b(ih, jh)$ if i or j is equal to 0 or N. Here, $U_{i,j}^k$ is an approximation of $U_{i,j}$ after k steps starting with an initial approximation U_{ij}^0 and α is a positive constant. It turns out that if α is sufficiently small, then the iteration converges, see Problem 73.9, although the convergence gets slower as the step size h decreases.

EXAMPLE 73.5. Assuming x_2 independence, we are led to the model

$$-u''(x) = f(x) \quad \text{for } 0 < x < 1, \quad u(0) = u_0,\, u(1) = u_1,$$

where $u'(x) = \frac{du}{dx}$. The corresponding discrete model takes the form

$$-(U_{i-1} - 2U_i + U_{i+1}) = h^2 f(ih),$$
$$i = 1, ..., N-1,\, U_0 = u_0,\, U_N = u_1, \quad (73.15)$$

with U_i representing an approximation of $u(ih)$. Assuming for simplicity $u_0 = u_1 = 0$, the discrete model can be written in the form

$$AU = b,$$

with $U = (U_1, ..., U_{N-1})$, $b = (b_1, ..., b_{N-1})$ with $b_i = h^2 f(ih)$, $A = (a_{ij})$ an $(N-1) \times (N-1)$ matrix with $a_{ii} = 2$, $a_{i,i-1} = a_{i-1,i} = -1$ and

$a_{ij} = 0$ if $|i-j| > 1$. The fixed point iteration described above can be written

$$U^{k+1} = U^k - \alpha(AU^k - b),$$

and the criterion of convergence is $\|I - \alpha A\| < 1$, which we prove in Problem 73.9 to be valid if $\alpha > 0$ is sufficiently small. Here, $\|I - \alpha A\|$ is the Euclidean norm of the matrix $I - \alpha A$ and the Spectral theorem implies

$$\|I - \alpha A\| = \max_i |1 - \alpha\lambda_i|,$$

where the λ_i are the eigenvalues of the symmetric matrix A.

73.8 The wave equation: vibrating elastic membrane

We now model the dynamic motion of the elastic membrane considered above in the static. We complement the given exterior force $f(x,t)$, which now may be dependent on time, by a dynamic force, which according to Newton's law, takes the form $m\ddot{u}$, with m representing mass per unit area and $\ddot{u}$ representing the acceleration of vertical displacement u. This to leads to the *wave equation*, modeling a vibrating membrane subject to an exterior load,

$$\begin{cases} \ddot{u} - \Delta u = f & \text{in } \Omega \times (0,T], \\ u = 0 & \text{on } \Gamma \times (0,T], \\ u(x,0) = u^0(x),\ \dot{u}(x,0) = \dot{u}^0(x) & \text{for } x \in \Omega, \end{cases} \tag{73.16}$$

where Ω denotes a domain in $\mathbb{R}^d$ with boundary Γ, u^0 is a given initial displacement, $\dot{u}^0$ is a given initial displacement velocity, and we assume homogeneous Dirichlet boundary conditions for simplicity. Other boundary conditions, notably periodic boundary conditions, are also relevant for this model.

73.9 Fluid mechanics

Fluid flow opens a rich field for mathematical modelling. We think of a fluid as a collection of many small "fluid particles" and we seek to describe the fluid flow resulting from the motion of all these fluid particles. We work under the assumption that the particles are so small and there are so many, that we can treat the fluid as a continuum. Usually, we use an *Eulerian* mode of description in which we describe the flow in terms of the *velocity* $u(x,t) \in \mathbb{R}^3$ of the fluid particles at position $x \in \mathbb{R}^3$ at time t, or simply the velocity of the fluid at (x, t). This corresponds to attaching an observer to each fixed point x for the purpose of observing the velocity $u(x, t)$ of the fluid particles that happen to be at position x at time t. The observer thus sits at position x and watches the fluid particles swirl by.

Alternatively, in a *Lagrangian* mode of description, an observer is attached to each fluid particle with the purpose of observing the change of velocity of the fluid particle with time. In this case, the observer follows the particle. The different modes of description are both useful and may also be used together, see the chapters on Convection-diffusion in CDE.

The equation of mass conservation

We consider the flow of a fluid within a certain volume $\Omega \in \mathbb{R}^3$ using an Eulerian description with $u(x, t)$ representing the velocity of the fluid at x at time t. The velocity u is a vector $u = (u_1, u_2, u_3)$.

Let $\rho(x, t)$ denote the *density* of a fluid at (x, t) measuring the mass of the fluid particles per unit of volume. Let V be a fixed volume with boundary S. The total mass of the fluid in V at time t is given by

$$\int_V \rho(x, t)\, dx.$$

The mass of fluid at time t passing out through the boundary S per unit of time is given by

$$\int_S \rho(x, t) u(x, t) \cdot n(x)\, ds(x) = \int_V \nabla \cdot (\rho u)(x, t)\, dx,$$

where we used the Divergence theorem. The rate of change of mass in V plus the rate of mass flow through the boundary must be zero if we assume that no fluid is added or removed, which leads to the following expression of *mass conservation*,

$$\frac{\partial}{\partial t} \int_V \rho(x, t)\, dx + \int_V \nabla \cdot (\rho u)(x, t)\, dx = 0.$$

If ρ varies smoothly, then $\frac{\partial}{\partial t}$ may be moved under the integral sign and since V was arbitrarily, we are led to the differential equation expressing *mass conservation*,

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho u) = 0, \tag{73.17}$$

Of course this is a basic equation of mathematical modelling. Performing the differentiation with respect to x, we can express mass conservation in the form

$$\frac{\partial \rho}{\partial t} + u \cdot \nabla \rho + \rho \nabla \cdot u = 0. \tag{73.18}$$

Particle paths and streamlines

Let the velocity of a fluid be given by the function $u(x,t)$. Consider the IVP

$$\frac{d}{dt}x(t) = u(x(t),t) \quad \text{for } t>0,\, x(0)=x_0.$$

The solution $x(t)$ represents the curve, or path or trajectory, followed by a fluid particle that starts at position x_0 at time $t=0$ and moves with velocity $u(x(t),t)$ for $t>0$. If the velocity $u(x,t)=u(x)$ is independent of time t, then particle paths are also referred to as *streamlines*.

Incompressible flow

If the fluid velocity $u(x,t)$ satisfies

$$\nabla\cdot u(x,t) = \Big(\frac{\partial u_1}{\partial x_1}+\frac{\partial u_2}{\partial x_2}+\frac{\partial u_3}{\partial x_3}\Big)(x,t)=0 \quad \text{for } x\in\Omega\, t>0,$$

then the flow is said to be *incompressible* in Ω for $t>0$.

If the flow is incompressible, the equation (73.18) of mass conservation takes the form

$$\frac{\partial\rho}{\partial t}+u\cdot\nabla\rho=0. \tag{73.19}$$

Since $\frac{dx}{dt}=u$ for a $x(t)$ particle path, the Chain rule implies

$$\frac{\partial}{\partial t}\rho(x(t),t)=\frac{\partial\rho}{\partial t}+u\cdot\nabla\rho=0.$$

This says that the density is constant along particle paths, or in other words the volume occupied by a certain set of fluid particles is constant. So, the fluid cannot be compressed. It is common to assume that the density of an incompressible fluid is constant.

Water is very nearly incompressible; it is very difficult to change the total volume of a bucket of water. Air is compressible; the air tank of a diver contains a huge volume of air at normal pressure compressed and stored in a small volume at high pressure. But to get the air into the tank consumes energy.

Incompressible potential flow

In so-called *stationary flow*, the velocity $u(x,t)$ is independent of time and thus the fluid velocity $u(x)$ is a function of $x\in\Omega$. Note that in a stationary flow the fluid particles at x are moving if $u(x)\neq 0$, but the velocity of the fluid particles at x does not change with time.

The velocity field $u(x)$ of *rotation-free* fluid flow satisfies $\nabla\times u=0$, which implies $u=\nabla\varphi$ for a scalar *velocity potential* φ under appropriate convexity assumptions. If the fluid is *incompressible*, then $\nabla \cdot u = 0$, and we obtain the Laplace equation $\Delta\varphi = 0$ for the potential of a rotation-free incompressible flow. At a solid boundary, through which the fluid cannot penetrate, the normal velocity of the fluid is zero, which translates into a homogeneous Neumann boundary condition $\partial_n\varphi = 0$ for the potential φ.

We now give some basic examples of incompressible potential flow. For simplicity, we consider situations in which the velocity $u(x)$ is independent of the x_3-coordinate.

EXAMPLE 73.6. The potential

$$\varphi(x_1, x_2) = x_1^2 - x_2^2$$

satisfies $\Delta\varphi = 0$ and the corresponding flow velocity $u = \nabla\varphi$ is given by

$$u(x) = (2x_1, -2x_2).$$

This represents stationary flow in a corner, see Fig. 73.5. A streamline $x(t)$ satisfies $\frac{dx}{dt} = (2x_1, -2x_2)$, which is a separable equation with solutions satisfying

$$x_1(t)x_2(t) = c,$$

where c is a constant, see Fig. 73.5. We check by computing $\frac{d}{dt}x_1x_2 = \dot{x}_1x_2 + x_1\dot{x}_2 = 2x_1x_2 - 2x_1 2x_2 = 0$.

EXAMPLE 73.7. The potential

$$\varphi(x) = \log(\|x\|)$$

satisfies $\Delta\varphi(x) = 0$ for $x \neq 0$, and the corresponding flow velocity $u = \nabla\varphi$ is given by $u(x) = \frac{x}{\|x\|^2}$, see Fig. 73.5.

EXAMPLE 73.8. We consider incompressible potential flow around an infinite circular cylinder along the x_3-axis with cross-section $\Omega = \{x = (x_1, x_2) \in \mathbb{R}^2 : \|x\| < 1\}$ from left to right, see Fig. 73.5. The potential φ is given in polar coordinates $x = r(\cos(\theta), \sin(\theta))$ by

$$\varphi(x) = \varphi(r, \theta) = (r + \frac{1}{r})\cos(\theta),$$

corresponding to a flow from right to left sweeping around Ω and approaching $u(x) = (1, 0)$ for $\|x_1\|$ large and x_2 bounded. We note that $\Delta\varphi = 0$ for $r \neq 0$, and that $\frac{\partial\varphi}{\partial r} = 1 - 1/r^2 = 0$ for $r = 1$ and thus the flow is tangential to the boundary of Ω.

Note that fluid flow is rarely rotation-free in the whole region occupied by the fluid. In particular, if the fluid is viscous then rotation is generated at solid boundaries.

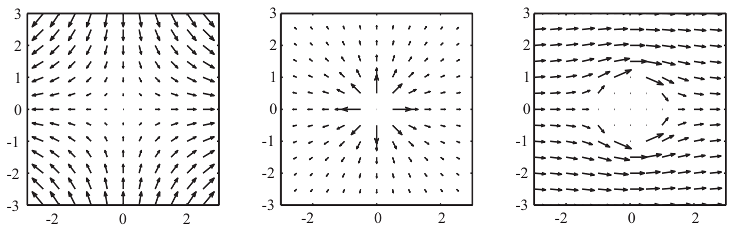


FIGURE 73.5. Examples of incompressible potential flow

Incompressible flow with rotation

We now consider basic examples of incompressible flow in two dimensions with non-zero rotation. We assume $u(x) = (u_1(x), u_2(x))$ satisfies $\nabla \cdot u = 0$, where $x = (x_1, x_2)$. Defining $v = (-u_2, u_1)$ this equation reads $\nabla \times v = 0$ and under appropriate convexity assumptions, there is a potential φ with $v = \nabla\varphi$. Thus,

$$u = (v_2, -v_1) = (\frac{\partial \varphi}{\partial x_2}, -\frac{\partial \varphi}{\partial x_1}) = \nabla \times \varphi.$$

With the rotation $\nabla \times u = f$ given, we are led to the Poisson equation for φ,

$$f = \nabla \times u = \nabla \times (\nabla \times \varphi) = -\Delta\varphi.$$

EXAMPLE 73.9. Given $f = 4$ we find the corresponding solution $\varphi(x) = -\|x\|^2$ with $u(x_1, x_2) = (-2x_2, 2x_1)$, see Fig. 73.6. Choosing $\varphi(x) = \log(\|x\|)$ corresponds to $f(x) = 0$ for $x \neq 0$, and the corresponding velocity $u(x_1, x_2) = \|x\|^{-2}(x_2, -x_1)$, see Fig. 73.6.

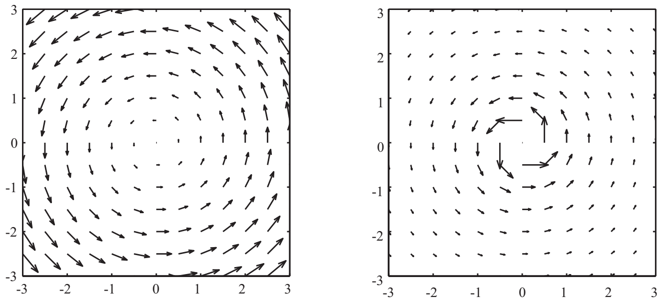


FIGURE 73.6. Incompressible flow with rotation

The Euler and Navier-Stokes equations

The *Euler equations* for an incompressible *inviscid* fluid with constant density equal to one, take the form

$$\frac{\partial u}{\partial t} + (u \cdot \nabla)u + \nabla p = f, \quad \nabla \cdot u = 0, \tag{73.20}$$

where $u(x,t)$ is the velocity and $p(x,t)$ the *pressure* of the fluid at the point x at time t, and f is an applied volume force like a gravitational force. In an inviscid fluid, the *viscosity* is zero and the only interior force acting between the fluid particles is the pressure force that is equal in all directions and acts normal to any surface. The equation $\nabla \cdot u = 0$ expresses the incompressibility of the flow. The first equation expresses Newton's law stating that the acceleration $\frac{d}{dt}u(x(t),t)$, where $x(t)$ is the trajectory followed by a fluid particle satisfying $\frac{dx}{dt} = u(x(t),t)$, is equal to the force $-\nabla p + f$, consisting of the pressure force $-\nabla p$ and the applied force f. We see this by computing with the the Chain rule and the equation $\frac{dx}{dt} = u(x(t),t)$ to get

$$\frac{d}{dt}u_i(x(t),t) = \frac{\partial u_i}{\partial t} + \frac{dx}{dt} \cdot \nabla u_i = \frac{\partial u_i}{\partial t} + (u \cdot \nabla)u_i,$$

which leads to the vector form (73.20). The *Navier-Stokes equations* are modifications of the Euler equations with an additional viscous force term $-\nu\Delta u$, where ν is the viscosity coefficient. In a fluid with non-zero viscosity, there are also tangential (shear) forces acting on a surface.

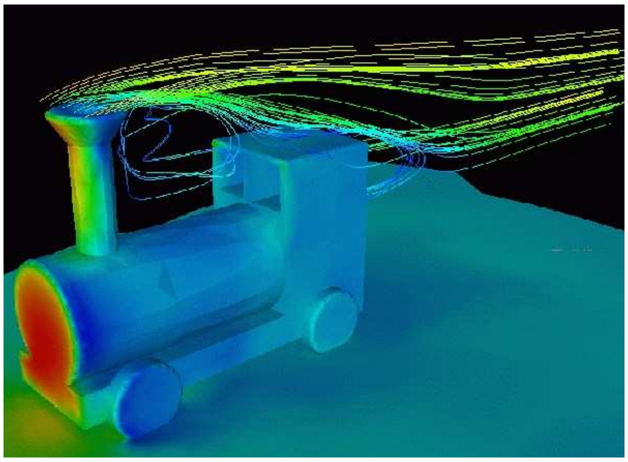


FIGURE 73.7. Flow of air and pressure around a toy train

73.10 Maxwell's equations

The interaction between electric and magnetic fields are described by *Maxwell's equations*:

$$\begin{cases} \dfrac{\partial B}{\partial t} + \nabla \times E = 0, \\ -\dfrac{\partial D}{\partial t} + \nabla \times H = J, \\ \nabla \cdot B = 0, \quad \nabla \cdot D = \rho, \\ B = \mu H, \quad D = \epsilon E, \quad J = \sigma E, \end{cases} \tag{73.21}$$

where E is the *electric field*, H is the *magnetic field* , D is the *electric displacement*, B is the *magnetic flux* , J is the *electric current*, ρ is the *charge*, μ is the *magnetic permeability* , ϵ is the *dielectric constant*, and σ is the *electric conductivity*. The first equation is referred to as *Faraday's law*, the second is *Ampère's law*, $\nabla \cdot D = \rho$ is *Coulomb's law*, *Gauss law* $\nabla \cdot B = 0$ expresses the absence of "magnetic charge", and $J = \sigma E$ is *Ohm's law*. Maxwell included the term $\partial D/\partial t$ for purely mathematical reasons and then used Calculus to predict the existence of electromagnetic waves before these had been observed experimentally.

FIGURE 73.8. Maxwell (1831-1879), inventor of the mathematical theory of electromagnetism: "We can scarcely avoid the conclusion that light consists in the transverse undulations of the same medium which is the cause of electric and magnetic phenomena".

Typical boundary conditions include various combinations of $E \cdot n$, $E \times n$, $H \cdot n$ and $H \times n$.

Maxwell's equations describe the whole world of electromagnetic phenomena with an astounding economy of notation and accuracy of modelling. Our modern information society builds on electromagnetic waves. We shall now pick out a couple of Laplace equation models from Maxwell's equations by considering some basic particular cases.

Electrostatics

A basic problem in *electrostatics* is to describe the stationary electric field $E(x)$ in a volume Ω in $\mathbb{R}^3$ containing *charges* of density $\rho(x)$ and enclosed by a perfectly conducting surface Γ. Faraday's law states that

$$\nabla \times E = 0 \quad \text{in } \Omega,$$

since we assume that $\frac{\partial B}{\partial t} = 0$. Recalling Chapter Potential fields, it follows that the electric field E is the gradient of a scalar *electric potential* φ, i.e. $E = \nabla\varphi$. Coulomb's law says

$$\nabla \cdot E = \rho \quad \text{in } \Omega,$$

so we are led to the Poisson equation for the potential φ,

$$\Delta\varphi = \nabla \cdot \nabla\varphi = \rho \quad \text{in } \Omega.$$

The boundary condition $E \times n = 0$ on the boundary Γ of Ω with n denoting the outward unit normal, says that the tangential component of E vanishes on the boundary. This models a perfectly conducting boundary in which differences in the electric field are leveled out. This means that $E = \nabla\varphi$ is normal to the boundary, so the boundary is a level surface of φ and the potential φ is constant on the boundary. Since φ is undetermined up to a constant, we may assume that $\varphi = 0$ on the boundary and we arrive at Poisson's equation $-\Delta\varphi = f$ with $f = -\rho$ in Ω with homogenous Dirichlet boundary conditions $\varphi = 0$ on Γ.

The potential $\varphi(x)$ of a point charge at the origin is given by

$$\varphi(x) = \frac{c}{\|x\|}$$

with the corresponding electric field

$$E(x) = -\frac{cx}{\|x\|^3}$$

and c a suitable constant. We shall return to this solution below.

EXAMPLE 73.10. Let $\Omega = \{x \in \mathbb{R}^2 : \|x\| < 1, x_1 < 0 \,\text{or}\, x_2 > 0\}$ be a circular disc with a piece cut out and a *reentrant* corner of angle $\omega = \frac{3\pi}{2}$, see Fig. 73.9. By a direct computation we can verify that the function

$$\varphi(x) = r^\alpha \sin(\alpha\theta)$$

expressed in polar coordinates $x = r(\cos(\theta), \sin(\theta))$, where $\alpha = \frac{\pi}{\omega} = \frac{2}{3}$, satisfies the Laplace equation $\Delta\varphi = 0$ in Ω and the boundary condition $\varphi = 0$ in the straight parts of the boundary meeting at the origin. Letting φ represent an electric potential, the corresponding electric field $E(x) = \nabla\varphi(x)$ satisfies

$$\frac{\partial E}{\partial r} = \alpha r^{\alpha-1}\sin(\alpha\theta)$$

and thus since $\alpha < 1$, is singular (infinite) at the corner where $r = 0$. This means that the electric field is very strong close to the corner, and the sharper the corner (α smaller) the stronger is the field. This may support the observation that an electric lightening is more likely to hit the pointed tower of church than a smooth hill, or the design of an electronic scanner where electrons pop out of the pin of a sharp needle.

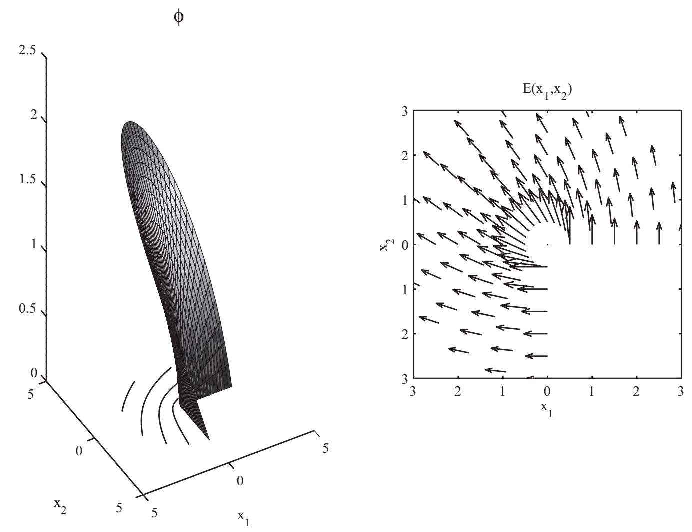


FIGURE 73.9. A potential with singular electric field

EXAMPLE 73.11. The potential φ of the electric field between two concentric spheres $S_1 = \{x \in \mathbb{R}^3 : \|x\| < r_1\}$ and $S_2 = \{x \in \mathbb{R}^3 : \|x\| < r_2\}$ with $r_2 > r_1$, is given by

$$\varphi(x) = \frac{1}{\|x\|}$$

if we assume that $\varphi = 1/r_i$ on S_i, $i = 1, 2$.

EXAMPLE 73.12. The function

$$\varphi(x_1, x_2)) = \arctan(\frac{x_2}{x_1})$$

defined for $x_1 > 0$ satisfies $\Delta\varphi(x) = 0$ for $x_1 > 0$, and is constant $= \arctan(c)$ on rays $x_2 = cx_1$ through the origin of slope c. The corresponding electric field $E(x) = \nabla\varphi(x)$ given by

$$E(x) = \frac{(-x_2, x_1)}{\|x\|^2}.$$

We see that $E(x)$ is singular at $x = 0$.

Magnetostatics

The basic problem in *magnetostatics* arises by combining Gauss' law $\nabla \cdot H = 0$, assuming μ constant and guaranteeing that $H = \nabla \times \varphi$ for some vector potential ψ satisfying $\nabla \cdot \psi = 0$, with Faraday's law $\nabla \times H = J$ to give

$$\nabla \times (\nabla \times \psi) = -\Delta\psi = J,$$

where we use the facts that $\nabla \times (\nabla \times \psi) = -\Delta\psi + \nabla(\nabla \cdot \psi)$ and that $\nabla \cdot \varphi = 0$.

The magnetic field around a unit current in the x_3-direction is given by

$$H(x) = \frac{1}{2\pi}\frac{(-x_2, x_1, 0)}{\|x\|^2},$$

which can be verified by direct computation showing that $\nabla \cdot H(x) = 0$ and $\nabla \times H(x) = 0$ for $(x_1, x_2) \neq 0$. The presence of the factor $\frac{1}{2\pi}$ corresponds to the unit current, which will be explained below in the Section Gravitation.

Time-dependent magnetics

In low frequency applications, the term $\frac{\partial D}{\partial t}$ so cleverly introduced by Maxwell, plays a minor role and can be discarded. Let's see where this leads. Since $\nabla \cdot B = 0$, we can write B as $B = \nabla \times \psi$, where ψ is a magnetic vector potential. Inserting this into Faraday's law gives

$$\nabla \times \left(\frac{\partial \psi}{\partial t} + E\right) = 0,$$

from which it follows that

$$\frac{\partial \psi}{\partial t} + E = \nabla\varphi,$$

for some scalar potential φ. Multiplying by σ and using the laws of Ohm and Ampère, we obtain a vector equation for the magnetic potential ψ:

$$\sigma\frac{\partial\psi}{\partial t} + \nabla\times\left(\mu^{-1}\nabla\times\psi\right) = \sigma\nabla\varphi.$$

This system reduces to a scalar equation in two variables if we assume that $B=(B_1,B_2,0)$ is independent of x_3. It then follows that ψ has the form $\psi=(0,0,u)$ for some scalar function u that depends only on x_1 and x_2, so that $B_1=\partial u/\partial x_2$ and $B_2=-\partial u/\partial x_1$. We end up with a scalar equation for the scalar magnetic potential u in the form

$$\sigma\frac{\partial u}{\partial t} - \nabla\cdot\left(\mu^{-1}\nabla u\right) = f, \tag{73.22}$$

for some function $f(x_1,x_2)$. Choosing $\sigma=\mu=1$ leads to the heat equation,

$$\begin{cases} \frac{\partial}{\partial t}u(x,t) - \Delta u(x,t) = f(x,t) & \text{for } x\in\Omega,\ 0<t\le T,\\ u(x,t)=0 & \text{for } x\in\Gamma,\ 0<t\le T,\\ u(x,0)=u_0(x) & \text{for } x\in\Omega, \end{cases} \tag{73.23}$$

where $\Omega\subset\mathbb{R}^2$ with boundary Γ, and we posed homogeneous Dirichlet boundary conditions. In the stationary case, we again obtain Poisson's equation with Dirichlet boundary conditions.

73.11 Gravitation

In his famous treatise *Mécanique Céleste* in five volumes published during 1799-1825, Laplace extended Newton's theory of gravitation and in particular developed a theory for describing gravitational fields based on using gravitational potentials that satisfy Laplace's equation, or more generally Poisson's equation.

We consider a gravitational field in $\mathbb{R}^3$ with gravitational force $F(x)$ at position x, generated by a distribution of mass of density $\rho(x)$. We recall that the work of a unit mass, moving along a curve Γ is given by

$$\int_\Gamma F\cdot ds,$$

If the curve Γ is closed, then the total work performed by the gravitational forces should be zero. Stokes' theorem implies that a gravitational field F should satisfy $\nabla\times F=0$. Using the basic result of the Chapter Potential fields, we conclude that F is the gradient of a scalar potential φ, i.e.

$$F(x)=\nabla\varphi(x). \tag{73.24}$$

Laplace proposed the following relation between the gravitational field F and the mass distribution ρ,

$$-\nabla \cdot F(x) = \rho(x), \tag{73.25}$$

assuming the gravitational constant is normalized to one. This is analogous to Coulomb's law $\nabla \cdot E = \rho$ in electrostatics and also to the energy balance equation $\nabla \cdot q = f$ for stationary heat conduction, where q is the heat flux and f a heat source. In particular, (73.25) states that $\nabla \cdot F(x) = 0$ at points x where there is no mass so that $\rho(x) = 0$. Combining (73.24) and (73.25), we obtain Poisson's equation $-\Delta\varphi = \rho$ for the gravitational potential φ.

Since the origin and property of gravitation of "acting at a distance" is still lacking a convincing physical explication, the equation $-\nabla \cdot F(x) = \rho(x)$ including $\nabla \cdot F = 0$ in empty space, should be viewed as a basic postulate on the nature of a gravitational field. Of course it seems very difficult to motivate that $\nabla \cdot F$ should be something different from zero in empty space, but a real a "proof" that $\nabla \cdot F$ must be zero in empty space seems to be missing.

Newton considered gravitational fields generated by *point masses*. Mathematically, a unit point mass at a point $z \in \mathbb{R}^3$ is represented by the so-called *delta function* δ_z at z, defined by the property that for any smooth function v,

$$\int_{\mathbb{R}^3} \delta_z \, v \, dx = v(z), \tag{73.26}$$

where the integration is to be interpreted in a generalized sense. We could think of a δ_z as a limit of positive functions $\varphi_h(x)$ such that $\varphi_h(x) = 0$ if $\|x - z\| > h$ and

$$\int_{\mathbb{R}^3} \varphi_h(x) \, dx = 1,$$

as h tends to zero. For example we may choose

$$\varphi_h(x) = \frac{3}{4\pi h^3} \quad \text{if } \|x - z\| < h$$

and $\varphi_h(x) = 0$ elsewhere. If $v(x)$ is Lipschitz continuous at z, then

$$lim_{h\to 0} \int_{\mathbb{R}^3} \varphi_h(x) v(x) \, dx = v(z),$$

which gives (73.26) its meaning. The function $\varphi_h(x)$ represents a very tall and narrow "hump" around z with volume one.

We expect that the gravitational potential $\Phi(x)$ corresponding to a unit point mass at the origin, to satisfy

$$-\Delta\Phi = \delta_0 \quad \text{in } \mathbb{R}^3, \tag{73.27}$$

assuming the gravitational constant to be equal to one. To give a precise meaning to this equation involving the somewhat mysterious delta function δ_0 at 0, we first formally multiply by a smooth *test function* v vanishing outside a bounded set to get

$$-\int_{\mathbb{R}^3} \Delta\Phi(x)v(x)\,dx = v(0). \tag{73.28}$$

Next, we integrate the left-hand side by parts formally using Green's formula to move the Laplacian from E to v, noting that the boundary terms disappear since v vanishes outside a bounded set. We may thus reformulate (73.27) as seeking a potential $E(x)$ satisfying

$$-\int_{\mathbb{R}^3} \Phi(x)\Delta v(x)\,dx = v(0), \tag{73.29}$$

for all smooth functions $v(x)$ vanishing outside a bounded set. We may view this as the concrete interpretation of (73.27), which is perfectly well defined since now the Laplacian acts on the smooth function $v(x)$ and the potential Φ is assumed to be integrable. We also require the potential $\Phi(x)$ to decay to zero as $\|x\|$ tends to infinity, which corresponds to a "zero Dirichlet boundary condition at infinity".

In the Chapter The divergence, rotation and Laplacian, we showed that the function $1/\|x\|$ satisfies Laplace's equation $\Delta u(x) = 0$ for $0 \neq x \in \mathbb{R}^3$, while it is singular at $x = 0$. We shall prove that the following scaled version of this function satisfies (73.29):

$$\Phi(x) = \frac{1}{4\pi}\frac{1}{\|x\|}. \tag{73.30}$$

We refer to this function as the *fundamental solution* of $-\Delta$ in $\mathbb{R}^3$. We conclude in particular that the gravitational field in $\mathbb{R}^3$ created by a unit point mass at the origin is given by

$$F(x) = \nabla\Phi(x) = -\frac{1}{4\pi}\frac{x}{\|x\|^3},$$

which is precisely Newton's inverse square law of gravitation. Laplace thus gives a motivation why the exponent should be two, which Newton did not (and therefore was criticized by Leibniz). Of course, it still remains to motivate (73.25). In the context of heat conduction, the fundamental solution $E(x)$ represents the stationary temperature in a homogeneous body with heat conductivity equal to one filling the whole of $\mathbb{R}^3$, subject to a concentrated heat source of strength one at the origin and with the temperature tending to zero as $\|x\|$ tends to infinity.

We now prove that the function $\Phi(x)$ defined by (73.30) satisfies (73.29). We first note that since Δv is smooth and vanishes outside a bounded set and $\Phi(x)$ is integrable over bounded domains,

$$\int_{\mathbb{R}^3} \Phi \Delta v \, dx = \lim_{a \to 0+} \int_{D_a} \Phi \Delta v \, dx, \tag{73.31}$$

where $D_a = \{x \in \mathbb{R}^3 : a < \|x\| < a^{-1}\}$, with a small positive, is a bounded region obtained from $\mathbb{R}^3$ by removing a little sphere of radius a with boundary surface S_a and also points further away from the origin than a^{-1}, see Fig. 73.10. We now use Green's formula on D_a with $w = \Phi$. Since

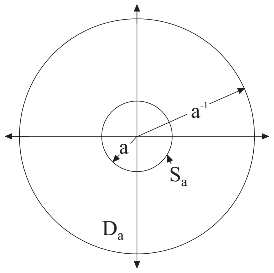


FIGURE 73.10. A cross-section of the domain D_a.

v is zero for $\|x\|$ large, the integrals over the outside boundary vanish when a is sufficiently small. Using the fact that $\Delta\Phi = 0$ in D_a, $\Phi = 1/(4\pi a)$ on S_a and $\partial\Phi/\partial n = 1/(4\pi a^2)$ on S_a with the normal pointing in the direction of the origin, we obtain

$$-\int_{D_a} \Phi \Delta v \, dx = \int_{S_a} \frac{1}{4\pi a^2} v \, ds - \int_{S_a} \frac{1}{4\pi a} \frac{\partial v}{\partial n} \, ds = I_1(a) + I_2(a),$$

with the obvious definitions of $I_1(a)$ and $I_2(a)$. Now, $\lim_{a\to 0} I_1(a) = v(0)$ because $v(x)$ is continuous at $x = 0$ and the surface area of S_a is equal to $4\pi a^2$, while $\lim_{a\to 0} I_2(a) = 0$. The desired equality (73.29) now follows recalling (73.31).

The corresponding fundamental solution of $-\Delta$ in $\mathbb{R}^2$ is given by

$$\Phi(x) = \frac{1}{2\pi} \log(\frac{1}{\|x\|}). \tag{73.32}$$

In this case the fundamental solution is not zero at infinity.

Replacing 0 by an arbitrary point $z \in \mathbb{R}^3$, (73.29) becomes

$$-\int_{\mathbb{R}^3} \Phi(z - x) \Delta v(x) \, dx = v(z), \tag{73.33}$$

which leads to a solution formula for Poisson's equation in $\mathbb{R}^3$. For example, if u satisfies the Poisson equation $-\Delta u = f$ in $\mathbb{R}^3$ and $|u(x)| = O(\|x\|^{-1})$ as

$\|x\| \to \infty$, then u may be represented in terms of the fundamental solution Φ and the right-hand side f as follows:

$$u(z) = \int_{\mathbb{R}^3} \Phi(z-x) f(x)\, dx = \frac{1}{4\pi} \int_{\mathbb{R}^3} \frac{f(x)}{\|z-x\|}\, dx. \tag{73.34}$$

We see that $u(z)$ is a mean value of f centered around z weighted so that the influence of the values of $f(x)$ is inversely proportional to the distance from z.

Similarly, the potential u resulting from a distribution of mass of density $\rho(x)$ on a (bounded) surface Γ in $\mathbb{R}^3$ is given by

$$u(z) = \frac{1}{4\pi} \int_{\Gamma} \frac{\rho(x)}{\|z-x\|}\, ds(x). \tag{73.35}$$

Formally, we obtain this formula by simply adding the potentials from all the different pieces of mass on Γ. One can show that the potential u defined by (73.35) is continuous in $\mathbb{R}^3$ if ρ is bounded on Γ, and of course u satisfies Laplace's equation away from Γ. Suppose now that we would like to determine the distribution of mass ρ on Γ so that the corresponding potential u defined by (73.35) is equal to a given potential u_0 on Γ, that is we seek in particular a function u solving the boundary value problem $\Delta u = 0$ in Ω and $u = u_0$ on Γ, where Ω is the volume enclosed by Γ. This leads to the following *integral equation*: given u_0 on Γ find the function ρ on Γ such that

$$\frac{1}{4\pi} \int_{\Gamma} \frac{\rho(y)}{\|x-y\|}\, ds(y) = u_0(x) \quad \text{for } x \in \Gamma. \tag{73.36}$$

This is a *Fredholm integral equation of the first kind*, named after the Swedish mathematician Ivar Fredholm (1866-1927). In the beginning of the 20th century, Fredholm and Hilbert were competing to prove the existence of solutions of the basic boundary value problems of mechanics and physics using integral equation methods. The integral equation (73.36) is an alternative way of formulating the boundary value problem of finding u such that $\Delta u = 0$ in Ω, and $u = u_0$ on Γ. Integral equations may also be solved using Galerkin methods. We return to the topic of integral equations and their numerical solution in the advanced volume.

73.12 Quantum Mechanics

The two most revolutionary achievements of physics during the 20th century was the development of of the *Theory of General Relativity* for Gravitation on astronomic scales by Einstein, and *Quantum Mechanics* for atomic scales by Schrödinger (1887-1961, Nobel Prize in Physics 1933). Einstein never fully accepted Quantum Mechanics, and the *Grand Unified Theory* connecting Gravitation and Quantum Mechanics is still missing, with *String Theory* being a recent attempt to fill the gap.

The basic equation of Quantum Mechanics is the *Schrödinger equation*, which for a system of N electrons (with the Born-Oppenheimer approximation) takes the following normalized form:

$$i\frac{\partial\varphi}{\partial t} - \sum_j \Delta_j\varphi + V(r_1, ..., r_N)\varphi = 0, \tag{73.37}$$

where $\varphi(r_1, ..., r_N, t)$ is a *wave function* depending on the set of space coordinates $(r_1, ..., r_N)$ with each r_j varying over $\mathbb{R}^3$, together with time t, Δ_j denotes the Laplacian with respect to the coordinate $r_j \in \mathbb{R}^3$, and $V(r_1, ..., r_N)$ denotes a *potential* depending on $(r_1, ..., r_N)$ representing repulsive Coulomb forces between the electrons and attractive Coulomb forces between the electrons and the (fixed) nuclei of the system, and i denotes the imaginary unit. The wave function is complex-valued and the square of its modulus represents an electron probability density.

The Schrödinger equation appears to give a very good description of phenomena on atomistic scales, but unfortunately it is not easy to deal with because of the large number of spatial dimensions involved: For a system with 100 electrons, which is still very small, the number of space dimensions is equal to 300, and standard techniques for either analytical or numerical solution fall very short. So, although the Schrödinger equation admittedly is a very beautiful equation which gives a surprisingly concise description of atomistic physics, it is certainly impossible to solve exactly analytically, and approximate solution becomes a key issue. The 1998 Nobel Prize in Chemistry was awarded Robert Kohn for his method for approximate solution of the Schrödinger equation based on using a single *electron density function* with the space dependence restricted to $\mathbb{R}^3$, independent of the number of electrons, and corresponding approximate potentials. Such simplified Schrödinger equations, referred to as *Kohn-Sham equations*, are today used extensively in computational chemistry.

The Hydrogen atom

The Hydrogen atom consisting of one electron and one neutron is the only case in which analytical solution of the Schrödinger equation is feasible: In this case the Schrödinger equation takes the following form assuming the neutron is positioned at the origin: Find the wave function $\varphi(x, t)$ with $x \in \mathbb{R}^3$, such that for $t > 0$

$$(i\frac{\partial}{\partial t} - \Delta + V)\varphi = 0 \quad \text{in } \mathbb{R}^3, \tag{73.38}$$

where Δ is the usual Laplacian with respect to x, $V(x) = -\frac{1}{|x|}$ is the Coulomb potential of the proton, with the normalization that

$$\int_{\mathbb{R}^3} |\varphi(x,t)|^2 \, dx = 1 \quad \text{for } t > 0.$$

For a domain $\Omega \in \mathbb{R}^3$, the integral

$$\int_{\Omega} |\varphi(x,t)|^2 \, dx$$

represents the probability to find the electron in the domain Ω at time t. In the time-harmonic case with a time-dependence of the form $exp(i\lambda)$ this leads to the eigenvalue problem: Find $\varphi(x) \neq 0$ and $\lambda \in \mathbb{R}$ such that

$$H\varphi = \lambda\varphi, \tag{73.39}$$

with $H = -\Delta + V$ the *Hamiltonian*, the potential $V(x) = -\frac{1}{|x|}$ and the eigenvalue λ representing the energy level. The eigenvalues are real and the (real) eigenfunction corresponding to the smallest eigenvalue (smallest energy) is referred to as the *ground state*.

Assuming spherical symmetry (73.39) takes the following form in spherical coordinates with r the radius: Find $\varphi(x) = \varphi(r)$ such that

$$-\frac{d^2\varphi}{dr^2} - \frac{2}{r}\frac{d\varphi}{dr} - \frac{2}{r}\varphi = \lambda\varphi \quad \text{for } r > 0,$$

with the side condition that $\varphi(0)$ is finite and $\varphi(x)$ is square integrable over $\mathbb{R}^3$. The ground state is given by the eigenfunction $\varphi(r) = \exp(-r)$ corresponding to the eigenvalue $\lambda = -1$.

Chapter 73 Problems

73.1. Interpret the fixed point iteration for Poisson's equation as an explicit time stepping scheme for the heat equation $\frac{du}{dt} - \Delta u = f$ with time step αh^2 with the starting value given by the initial approximation U^0. Explain why the convergence is slow if h is small.

73.2. Consider a horizontal elastic membrane spanned over a circular ring with constant tension H in all directions in unloaded configuration. Discuss under what conditions the membrane can carry a non-zero volume of water and try to compute the volume.

73.3. Prove that (73.32) is a fundamental solution of $-\Delta$ in $\mathbb{R}^2$.

73.4. Because the presented mathematical models of heat flow and gravitation, namely Poisson's equation, are the same, it opens the possibility of thinking of a gravitational potential as "temperature" and a gravitational field as "heat flux". Can you "understand" something about gravitation using this analogy?

73.5. Present the integral equation corresponding to (73.36) in the case $d = 2$.

73.6. What equation is obtained if $\partial D/\partial t$ is not neglected in the setting of time-dependent magnetics, but the x_3 independence is kept?

73.7. Derive the heat equation describing the heat conduction in a thin piece of wire of length one whose ends are kept at a fixed temperature (i.e., derive the heat equation in one dimension):

$$\begin{cases} \dot{u} - u'' = f & \text{in } (0,1) \times (0,T], \\ u(0,t) = u(1,t) = 0 & \text{for } t \in (0,T], \\ u(x,0) = u_0(x) & \text{for } x \in (0,1). \end{cases} \tag{73.40}$$

73.8. Let $F(x)$ be the gravitational field generated by a homogeneous ball of mass m occupying the volume $\{x \in \mathbb{R}^3 : \|x\| \leq r\}$, satisfying $\nabla F(x) = \rho$ for $\|x\| < r$ and $\nabla F(x) = 0$ for $\|x\| > r$, where ρ is the density of the sphere. Argue that by symmetry $F(x) = f(\|x\|)\frac{-x}{\|x\|}$ for $\|x\| > r$ for some function $f : (0, \infty) \to \mathbb{R}$. Use the Divergence theorem to see that if $R > r$ then

$$\int_{S_R} F(x) \cdot n dS = 4\pi R^2 f(R) = \int_{B_R} \nabla F(x)\, dx = m,$$

where S_R is the boundary of the ball $B_R = \{x \in \mathbb{R}^3 : \|x\| \leq R\}$. Conclude that $f(R) = \frac{m}{4\pi R^2}$, and thus that $F(x) = \frac{m}{4\pi}\frac{-x}{\|x\|^3}$ for $\|x\| > r$. This gives an alternative way of handling of Newton's nightmare. Note the change of normalization with the factor $1/4\pi$ appearing here.

73.9. To analyze the convergence of the fixed point iteration for the system of equations (73.15), we need to show that $\|I - \alpha A\| < 1$, where $A = (a_{ij})$ is the $(N-1) \times (N-1)$ matrix with $a_{ii} = 2$, $a_{i,i-1} = a_{i-1,i} = -1$ and $a_{ij} = 0$ if $|i-j| > 1$. Since A is symmetric, we have recalling the Chapter The Spectral Theorem:

$$\|I - \alpha A\| = \max_i |1 - \alpha \lambda_i|,$$

where λ_i, $i = 1, ..., N-1$, are the eigenvalues of A. To see this, diagonalize. Prove that for all nonzero $V \in \mathbb{R}^{N-1}$

$$AV \cdot V = \sum_{i,j=1}^{N-1} a_{ij} V_i V_j > 0,$$

and conclude that $\lambda_i > 0$ for all i (Hint: complete squares!). Show similarly that for all $V \in \mathbb{R}^{N-1}$

$$(I - \alpha A)V \cdot V \geq 0$$

if $\alpha \leq \frac{1}{4}$ (Hint: same as before!). Conclude that Fixed point iteration converges if $0 < \alpha \leq \frac{1}{4}$. Can you prove convergence if $\alpha < \frac{1}{2}$?. What about convergence if $\alpha < 0$?. Hint: use that if A is a symmetric $m \times m$ matrix with eigenvalues $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_m$, then $\lambda_1 = \min_{V \in \mathbb{R}^m}(AV \cdot V)/(V \cdot V)$ and $\lambda_m = \max_{V \in \mathbb{R}^m}(AV \cdot V)/(V \cdot V)$, where $V \neq 0$.

73.10. Extend the above analysis to the 5-point scheme for the Laplacian and show that fixed point iteration converges if $0 < \alpha < \frac{1}{8}$ (or better $\alpha < \frac{1}{4}$).

73.11. Gather some friends and arrange them in a square regular grid, and ask them to keep updating their own value according to a Svensson's formula as the mean value of their neighbors (starting with zero), and assigning certain given values to the people at the boundary. Collect the values obtained after convergence. You have solved Laplace equation on a square with Dirichlet boundary values numerically. What value of α in fixed point iteration did you effectively use?

73.12. Prove Bernoulli's theorem stating that in stationary Euler flow satisfying $(u \cdot \nabla)u + \nabla p = 0$ the quantity $\frac{1}{2}\|u|^2 + p$ is constant along streamlines.

73.13. Explain the *Magnus effect* causing a top-spin tennis ball curve downwards.

> The idea of the continuum seems simple to us. We have somehow lost sight of the difficulties it implies ... We are told such a number as the square root of 2 worried Pythagoras and his school almost to exhaustion. Being used to such queer numbers from early childhood, we must be careful not to form a low idea of the mathematical intuition of these ancient sages; their worry was highly credible. (Schrödinger)

74
Chemical reactions*

We already know the laws that govern the behavior of matter under all but the most extreme situations. In particular, we know the basic laws that underlie all of chemistry and biology. Yet we have certainly not reduced these objects to the status of solved problems; we have, as yet, had little success in predicting human behavior from mathematical equations. So even if we do find a complete set of basic laws, there will still be in the years ahead the intellectual challenging task of developing better approximation methods, so that we can make useful predictions of the probable outcomes in complicated and realistic situations. (S. Hawking in A Brief History of Time)

It is especially difficult to find exact solutions of the equations, as the equations (Einstein's equations) are non-linear. (Einstein)

Inasmuch as a propagating flame may be considered as a wave of chemical reactions sweeping across a flowing gas, it offers an excellent proving ground for the analytical skills of a fluid dynamicist, a heat and mass transfer specialist and a physical chemist, all put together into a well-rounded applied mathematician. (M. Kanury)

74.1 Constant temperature

We consider N different chemical species A_1,..., A_N, which participate in J reactions with stoichiometric (positive) integer coefficients $\nu_{n,j}$ for species n appearing as reactant in reaction j and $\lambda_{n,j}$ for species n appearing as product in reaction j (with the coefficients being zero if the species is not a reactant or product). This is commonly expressed as

$$\sum_{n=1}^{N} \nu_{n,j} A_n \to \sum_{n=1}^{N} \lambda_{n,j} A_n \quad \text{for } j = 1, ..., J. \tag{74.1}$$

We say that the *order* of reaction j is equal to $\sum_{n=1}^{N} \nu_{n,j}$. We denote the *molar concentration* (expressed in moles per unit volume) of species A_n by c_n. The *reaction rate* r_j of reaction j is supposed to be given by

$$r_j = k_j(T) \prod_{m=1}^{N} c_m^{\nu_{m,j}},$$

where the *reaction coefficient* or *Arrhenius factor* $k_j(T)$ is given by

$$k_j(T) = B_j T^{\alpha_j} \exp(-\frac{E_j}{RT}),$$

with $E_j > 0$ representing the *activation energy*, $B_j T^{\alpha_j}$ representing the *frequency factor*, B_j and α_j are positive constants, the absolute temperature T is assumed to be the same for all species, and R is the gas constant. The basic idea behind the product formula $\prod_{m=1}^{N} c_m^{\nu_{m,j}}$ is that the reaction rate is proportional to the molar concentrations of the reactants with each reactant A_m counted $\nu_{m,j}$ times. The Arrhenius factor is small if T is below some threshold value corresponding to the quotient $\frac{E_j}{RT}$ being moderately large.

The net production rate (moles per volume per unit time) of species A_n in reaction j is given by $\alpha_{n,j} r_j$, where

$$\alpha_{n,j} = \lambda_{n,j} - \nu_{n,j},$$

and the total net production rate s_n of species n is given by

$$s_n = \sum_{j=1}^{J} \alpha_{n,j} r_j.$$

We now assume that the temperature T is constant and is given, and we seek the vector of concentration $c(t) = (c_1(t), ..., c_N(t))$ as a function of time t describing the dynamics of the set of reactions for $t > 0$, assuming that $c(0) = c^0$, where $c^0 = (c_1^0, ..., c_N^0)$ is a given vector of initial concentrations. Using the balance equation $\dot{c}_n = s_n$ for each species $n = 1, 2, ..., N$, we obtain the following initial value problem for a system of ordinary differential equations: Find $c(t) = (c_1(t), ..., c_N(t))$ such that

$$\begin{cases} \dot{c}_n(t) = \sum_{j=1}^{J} \alpha_{n,j} k_j(T) \prod_{m=1}^{N} c_m(t)^{\nu_{m,j}} \quad \text{for } t > 0,\ n = 1, ..., N, \\ c(0) = c^0. \end{cases} \tag{74.2}$$

This is an initial value problem of the form $\dot{u}(t) = f(u(t))$ for $t > 0$, $u(0) = u^0$, where $u(t) = c(t)$ and $f : \mathbb{R}^N \to \mathbb{R}^N$ is a given function.

An *Equilibrium* for a given temperature T corresponding to $\dot{c}_n(t) = 0$ for $t > 0$, $n = 1, ..., N$, is characterized by the algebraic system of equations

$$\sum_{j=1}^{J} \alpha_{n,j} k_j(T) \prod_{m=1}^{N} c_m^{\nu_{m,j}} = 0 \quad n = 1, ..., N, \tag{74.3}$$

corresponding to the equation $f(u) = 0$.

Example 74.1. The reaction

$$2\mathrm{NO} + \mathrm{Cl}_2 \to 2\mathrm{NOCl},$$

can be put in the form (74.1) with $A_1 = \mathrm{NO}$, $A_2 = \mathrm{Cl}_2$, $A_3 = \mathrm{NOCl}$, $N = 3$, $J = 1$, $\nu_{1,1} = 2$, $\nu_{2,1} = 1$, $\nu_{3,1} = 0$, $\lambda_{1,1} = 0$, $\lambda_{2,1} = 0$, $\lambda_{3,1} = 2$, $\alpha_{1,1} = -2$, $\alpha_{2,1} = -1$, and $\alpha_{3,1} = 2$.

Example 74.2. The two reactions

$$2\mathrm{NO} + \mathrm{Cl}_2 \to^{k_1} 2\mathrm{NOCl},$$
$$2\mathrm{NOCl} \to^{k_2} 2\mathrm{NO} + \mathrm{Cl}_2,$$

can be put in the form (74.1) with $A_1 = \mathrm{NO}$, $A_2 = \mathrm{Cl}_2$, $A_3 = \mathrm{NOCl}$, $N = 3$, $J = 2$, $\nu_{1,1} = 2$, $\nu_{2,1} = 1$, $\nu_{3,1} = 0$, $\lambda_{1,1} = 0$, $\lambda_{2,1} = 0$, $\lambda_{3,1} = 2$, $\alpha_{1,1} = -2$, $\alpha_{2,1} = -1$, $\alpha_{3,1} = 2$, $\nu_{1,2} = 0$, $\nu_{2,2} = 0$, $\nu_{3,2} = 2$, $\lambda_{1,2} = 2$, $\lambda_{2,2} = 1$, $\lambda_{3,2} = 0$, $\alpha_{1,2} = 2$, $\alpha_{2,2} = 1$, and $\alpha_{3,2} = -2$. Equilibrium is characterized by

$$k_1 c_1^2 c_2 = k_2 c_3^2, \quad \text{or} \quad \frac{c_1^2 c_2}{c_3^2} = \frac{k_2}{k_1}.$$

Example 74.3. An *ideal first order tank reactor* is modeled by the equation

$$qc^0 - Vkc = qc,$$

where c^0 is the reactant concentration at inflow, c is the concentration in the reactor, q is the inflow (= outflow) rate, V is the volume of the reactor and k is a reaction coefficient. The equation expresses that the (rate of) reactant inflow minus the reactant consumed in the reaction is equal to the reactant outflow. Introducing $\tau = \frac{V}{q}$, which is the time the reactant stays in the reactor, we get

$$c = \frac{c^0}{1 + \tau k}.$$

The *efficiency* of the reactor is given by

$$\eta = \frac{c^0 - c}{c^0} = \frac{\tau k}{1 + \tau k} = \frac{1}{1 + \frac{1}{\tau k}}.$$

We see in particular that the efficiency decreases as τ decreases.

EXAMPLE 74.4. An *ideal first order tube reactor* occupying the interval $(0,1)$, which may be viewed as a set of ideal first order tank reactors coupled in series, is modeled by

$$qc(x) - A\Delta x k c(x) = qc(x+\Delta x) \quad \text{for } 0 < x < 1,$$

where q is the (constant) flow rate, A the cross section of the tube, and Δx is a small increment in x. Dividing by Δx and letting Δx tend to zero leads to the initial value problem of finding the concentration $c(x)$ for $0 \le x \le 1$ such that

$$\frac{dc}{dx} = -\tau kc \quad \text{for } 0 < x \le 1, \quad c(0) = c^0,$$

where $\tau = \frac{A}{q}$. The solution is given by

$$c(x) = c^0 e^{-\tau kx},$$

and the efficiency $\eta = \frac{c^0 - c(1)}{c^0}$ is

$$\eta = 1 - e^{-\tau k}.$$

Using the fact that $\frac{x}{1+x} < 1 - e^{-x}$ for $x > 0$, it follows that the ideal tube reactor is more efficient than the ideal tank reactor.

74.2 Variable temperature

Suppose now that the temperature $T(t)$ is variable with time t, and is unknown along with the concentrations $c_1(t), ..., c_N(t)$. The *heat of reaction* of reaction j is given by

$$(-\sum_{m=1}^{N} \alpha_{m,j} h_m) r_j,$$

where h_m is the *molar enthalpy* of species A_m. The heat of reaction is positive for an *exothermic reaction* and negative for an *endothermic reaction.*

The problem is now to find $c(t) = (c_1(t), ..., c_N(t))$ and $T(t)$ for $t > 0$ such that

$$\begin{cases} \dot{c}_n = \sum_{j=1}^{J} \alpha_{n,j} k_j(T) \prod_{m=1}^{N} c_m^{\nu_{m,j}}, & t > 0,\, n = 1, ..., N, \\ C_p \dot{T} = \sum_{j=1}^{J} (-\sum_{m=1}^{N} \alpha_{m,j} h_m) k_j(T) \prod_{m=1}^{N} c_j^{\nu_{m,j}}, & \\ c(0) = c^0,\, T(0) = T^0, & \end{cases} \tag{74.4}$$

where $c^0 = (c_1^0, ..., c_N^0)$ and T^0 are given initial concentrations and temperature, and C_p is the specific heat of the mixture of species.

74.3 Space dependence

Adding spacial dependence in a domain Ω in $\mathbb{R}^3$, we are led to the following model: Find $c(x,t) = (c_1(x,t), ..., c_N(x,t))$ and $T(x,t)$ for $x \in \Omega$, $t > 0$, such that

$$\begin{cases} \dot{c}_n + \nabla \cdot (c_n \beta) - \nabla \cdot (\epsilon_n \nabla c_n) \\ \qquad = \sum_{j=1}^{J} \alpha_{n,j} k_j(T) \prod_{m=1}^{N} c_m^{\nu_{m,j}} \quad \text{for } x \in \Omega,\, t > 0,\, n = 1, ..., N, \\ C_p \dot{T} + \nabla \cdot (C_p T \beta) - \nabla \cdot (\epsilon_0 \nabla T) \\ \qquad = \sum_{j=1}^{J} (-\sum_{m=1}^{N} \alpha_{m,j} h_m) k_j(T) \prod_{m=1}^{N} c_m^{\nu_{m,j}} \quad \text{for } x \in \Omega,\, t > 0, \\ c(x,0) = c^0, \quad T(x,0) = T^0 \quad \text{for } x \in \Omega, \end{cases} \tag{74.5}$$

where $\beta(x,t)$ is a given convection velocity, and the ϵ_n are given diffusion coefficients. The system is complemented by boundary conditions of Dirichlet, Neumann or Robin type for each equation.

EXAMPLE 74.5. A stationary one species constant temperature first order reaction with constant diffusion and zero convection is modeled in dimensionless form by the equation

$$\Delta u = \varphi^2 u \quad \text{in } \Omega,$$

together with Dirichlet, Neumann or Robin boundary conditions, where φ is the *Thiele modulus*, and Ω is a domain in $\mathbb{R}^d$, $d = 1, 2, 3$. A quantity of interest as a function of Ω, the reaction coefficient φ^2 and the boundary conditions, is the *total production*

$$\int_\Omega u(x)\, dx.$$

EXAMPLE 74.6. A simple model for *flame propagation* in a channel takes the form

$$\begin{cases} \dot{u}_1 - \Delta u_1 + \beta_1 \frac{\partial u_1}{\partial x_1} = u_2 f(u_1) \qquad x \in \Omega,\, t > 0, \\ \dot{u}_2 - \Delta u_2 + \beta_1 \frac{\partial u_2}{\partial x_1} = -u_2 f(u_1) \quad x \in \Omega,\, t > 0,\, x \in \Omega, \end{cases} \tag{74.6}$$

together with appropriate boundary conditions, where $\Omega = \mathbb{R} \times (0,1)$, u_1 represents temperature, u_2 represents a reactant concentration, β_1 is the velocity of the reactant in the x_1 direction, and $u_2 f(u_1)$ represents a reaction rate with $f : \mathbb{R}^+ \to \mathbb{R}^+$ given. With a proper choice of β_1 we may seek a stationary solution with $\dot{u} = 0$ corresponding to a propagating flame front.

EXAMPLE 74.7. A basic model for *combustion* in a domain Ω in $\mathbb{R}^3$ takes the form: Find the concentration c and temperature T such that:

$$\begin{cases} \dot{c} - \epsilon_1 \Delta c = -B_1 e^{-\frac{E}{RT}} c, \quad x \in \Omega,\, t > 0, \\ \dot{T} - \epsilon_0 \Delta T = B_0 e^{-\frac{E}{RT}} c \qquad x \in \Omega,\, t > 0, \end{cases} \tag{74.7}$$

together with, say, homogeneous Neumann boundary conditions, and with B_0 and B_1 positive constants. Depending on the activation energy E and initial conditions, the process may be fast or slow locally in space and time.

75
Calculus Tool Bag II

Axiom 1: All bodies are either in motion or at rest.
Axiom 2: Each single body can move at varying speeds.
Lemma 1: Bodies are distinguished from one another in respect of motion and rest, quickness and slowness, and not in respect of substance.
Lemma 2: All bodies agree in certain respects.
Lemma 3: A body in motion or at rest must have been determined to motion or rest by another body, which likewise has been determined to motion or rest by another body, and that body by another, and so ad infinitum.
...
Lemma 6: If certain bodies composing an individual thing are made to change the existing direction of their motion, but in such a way that they can continue their motion and keep the same mutual relation as before, the individual thing will likewise preserve the same mutual relation as before, the individual thing will likewise preserve its own nature without change of form. (Spinoza 1632-1677, Ethica II)

Timeo hominem unius libri. (St. Thomas of Aquino)

75.1 Introduction

We here collect the basic tools of Calculus of functions $f : \mathbb{R}^n \to \mathbb{R}^m$, that is Calculus of vector-valued functions of several real variables. The Euclidean norm of a vector $x = (x_1, ..., x_n) \in \mathbb{R}^n$ is denoted by $\|x\| = \sum_{i=1}^n x_i^2$.

75.2 Lipshitz continuity

A function $f : A \to \mathbb{R}^m$ with a subset of $\mathbb{R}^n$ is Lipschitz continuous on A if there is a constant L such that

$$\|f(x) - f(y)\| \leq L\|x - y\| \quad \text{for all } x, y \in A.$$

75.3 Differentiability

A function $f : A \to \mathbb{R}^m$ is *differentiable at* $\bar{x} \in A$, where A is an open subset of $\mathbb{R}^n$, if there is a $m \times n$ matrix $f'(\bar{x})$, called the *Jacobian* of the function $f(x)$ at $\bar{x}$, and a constant $K_f(\bar{x})$, such that for all $x \in A$ close to $\bar{x}$,

$$f(x) = f(\bar{x}) + f'(\bar{x})(x - \bar{x}) + E_f(x, \bar{x}),$$

where $E_f(x, \bar{x})$ is an m-vector satisfying $\|E_f(x, \bar{x})\| \leq K_f(\bar{x})\|x - \bar{x}\|^2$. We say that $f : A \to \mathbb{R}^m$ is *uniformly differentiable on* A if the constant $K_f(\bar{x}) = K_f$ can be chosen independently of $\bar{x} \in A$. We write $f' = \nabla f$ if $m = 1$ and call ∇f the gradient of f.

75.4 The Chain rule

If $g : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable at $\bar{x} \in \mathbb{R}^n$, and $f : \mathbb{R}^m \to \mathbb{R}^p$ is differentiable at $g(\bar{x}) \in \mathbb{R}^m$ and further $g : \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous, then the composite function $f \circ g : \mathbb{R}^n \to \mathbb{R}^p$ is differentiable at $\bar{x} \in \mathbb{R}^n$ with Jacobian

$$(f \circ g)'(\bar{x}) = f'(g(\bar{x}))g'(\bar{x}).$$

75.5 Mean Value theorem for $f : \mathbb{R}^n \to \mathbb{R}$

If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable on $\mathbb{R}^n$ with a Lipschitz continuous gradient ∇f, then for given x and $\bar{x}$ in $\mathbb{R}^n$, there is $y = x + \bar{t}(x - \bar{x})$ with $\bar{t} \in [0, 1]$, such that

$$f(x) - f(\bar{x}) = \nabla f(y) \cdot (x - \bar{x}).$$

75.6 A minimum point is a stationary point

If $\bar{x} \in \mathbb{R}^n$ is a *local minimum point* of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, that is, $f(\bar{x}) \leq f(x)$ for all x cloes to $\bar{x}$, then $\nabla f(\bar{x}) = 0$.

75.7 Taylor's theorem

If $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable with Lipschitz continuous Hessian $H = (h_{ij})$ with elements $h_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$, then, for given x and $\bar{x} \in \mathbb{R}^n$, there is $y = x + \bar{t}(x - \bar{x})$ with $\bar{t} \in [0, 1]$, such that

$$\begin{aligned} f(x) &= f(\bar{x}) + \nabla f(\bar{x}) \cdot (x - \bar{x}) + \frac{1}{2} \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j}(y)(x_i - \bar{x}_i)(x_j - \bar{x}_j) \\ &= f(\bar{x}) + \nabla f(\bar{x}) \cdot (x - \bar{x}) + \frac{1}{2}(x - \bar{x})^\top H(y)(x - \bar{x}). \end{aligned}$$

75.8 Contraction mapping theorem

If $g : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous with Lipschitz constant $L < 1$, then the equation $x = g(x)$ has a unique solution $\bar{x} = \lim_{i \to \infty} x^{(i)}$, where $\{x^{(i)}\}_{i=1}^{\infty}$ is a sequence in $\mathbb{R}^n$ generated by Fixed Point Iteration: $x^{(i)} = g(x^{(i-1)})$, starting with any initial value $x^{(0)}$.

75.9 Inverse Function theorem

Let $f : \mathbb{R}^n \to \mathbb{R}^n$ and assume the coefficients of $f'(x)$ are Lipschitz continuous close to $\bar{x}$ and $f'(\bar{x})$ is non-singular. Then for y sufficiently close to $\bar{y} = f(\bar{x})$, the equation $f(x) = y$ has a unique solution x. This defines x as a function $x = f^{-1}(y)$ of y.

75.10 Implicit Function theorem

If $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ with $f(x, y) \in \mathbb{R}^n$ and $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, $f(\bar{x}, \bar{y}) = 0$, and the Jacobian $f'_x(x, y)$ with respect to x is Lipschitz continuous for x close to $\bar{x}$ and y close to $\bar{y}$, and $f'_x(\bar{x}, \bar{y})$ is non-singular, then for y close to $\bar{y}$, the equation $f(x, y) = 0$ has a unique solution $x = g(y)$, which defines x as a function $g(y)$ of y.

75.11 Newton's method

If $\bar{x}$ is a root of $f : \mathbb{R}^n \to \mathbb{R}^n$ such that $f(x)$ is uniformly differentiable with a Lipschitz continuous derivative close to $\bar{x}$ and $f'(\bar{x})$ is non-singular, then Newton's method $x^{(i+1)} = x^{(i)} - f'(x^{(i)})^{-1} f(x^{(i)})$ for solving $f(x) = 0$ converges quadratically if started sufficiently close to $\bar{x}$.

75.12 Differential operators

Gradient of a function $u : \mathbb{R}^d \to \mathbb{R}$:

$$\text{grad } u = \nabla u = \left(\frac{\partial u}{\partial x_1}, \frac{\partial u}{\partial x_2}, ..., \frac{\partial u}{\partial x_d}\right).$$

Divergence of a vector function $u : \mathbb{R}^d \to \mathbb{R}^d$:

$$\text{div } u = \nabla \cdot u = \sum_{i=1}^{d} \frac{\partial u_i}{\partial x_i}.$$

Rotation of a vector function $u : \mathbb{R}^3 \to \mathbb{R}^3$:

$$\text{rot } u = \nabla \times u = \left(\frac{\partial u_3}{\partial x_2} - \frac{\partial u_2}{\partial x_3}, \frac{\partial u_1}{\partial x_3} - \frac{\partial u_3}{\partial x_1}, \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}\right).$$

Laplacian of a function $u : \mathbb{R}^d \to \mathbb{R}$:

$$\Delta u = \nabla \cdot (\nabla u) = \text{div (grad } u) = \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2}.$$

Identities:

$$\begin{aligned} \nabla \cdot (\nabla \times u) &= 0, \\ \nabla \times (\nabla u) &= 0, \\ \nabla \times (\nabla \times u) &= -\Delta u + \nabla(\nabla \cdot u). \end{aligned}$$

Laplacian in $\mathbb{R}^2$ in polar coordinates $x = (x_1, x_2) = (r\cos(\theta), r\sin(\theta))$:

$$\Delta u = \frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}.$$

Laplacian in spherical coordinates $x = (r\sin(\varphi)\cos(\theta), r\sin(\varphi)\sin(\theta), r\cos(\varphi))$:

$$\Delta u = \frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right) + \frac{1}{r^2\sin(\theta)}\frac{\partial}{\partial \theta}\left(\sin(\theta)\frac{\partial u}{\partial \theta}\right) + \frac{1}{r^2\sin^2(\theta)}\frac{\partial^2 u}{\partial \varphi^2}.$$

The Laplacian is invariant under orthogonal coordinate transformations in $\mathbb{R}^d$.

75.13 Curve integrals

If $\Gamma = s([a,b])$ is a curve in $\mathbb{R}^n$ given by the function $s : [a,b] \to \mathbb{R}^n$, and $u : \Gamma \to \mathbb{R}$, then

$$\begin{aligned}\int_\Gamma u\, ds = \int_\Gamma u(x)\, ds(x) &\equiv \int_a^b u(s(t))\|s'(t)\|\, dt,\\ \int_\Gamma u \cdot ds = \int_a^b u(s(t)) \cdot s'(t)\, dt,&\\ \int_\Gamma ds = \int_a^b \|s'(t)\|\, dt = \text{length of } \Gamma.&\end{aligned}$$

If $u = \nabla\varphi$, then

$$\int_\Gamma u \cdot ds = \varphi(s(b)) - \varphi(s(a)).$$

75.14 Multiple integrals

Integral over the unit square: If $f : Q = [0,1] \times [0,1] \to \mathbb{R}$ is Lipschitz continuous, then

$$\int_Q f(x)\, dx = \int_0^1 \int_0^1 f(x_1, x_2)\, dx_1\, dx_2 = \lim_{n\to\infty} \sum_{i=1}^N \sum_{j=1}^N f(x_{1,i}^n, x_{2,j}^n) h_n h_n,$$

where $h_n = 2^{-n}$, $x_{j,i}^n = ih_n$, $N = 2^n$, and

$$\int_Q f(x)\, dx == \int_0^1 \Big(\int_0^1 f(x_1, x_2)\, dx_2\Big)\, dx_1 = \int_0^1 \Big(\int_0^1 f(x_1, x_2)\, dx_1\Big)\, dx_2.$$

Change of variables: If $y \to x = g(y)$ maps a domain $\tilde{\Omega}$ in $\mathbb{R}^d$ onto a domain Ω in $\mathbb{R}^d$, where the Jacobian of g is Lipschitz continuous and $f : \Omega \to \mathbb{R}$ be Lipschitz continuous, then

$$\int_\Omega f(x)\, dx = \int_{\tilde{\Omega}} f(g(y))|\det g'(y)|\, dy,$$

Polar coordinates:

$$\int_\Omega f(x_1, x_2)\, dx_1 dx_2 = \int_{\tilde{\Omega}} f(r\cos(\theta, r\sin(\theta))\, r dr\, d\theta,$$

where $(r, \theta) \to x$ is a one-to-one mapping of $\tilde{\Omega}$ onto Ω given by $x = (r\cos(\theta, \sin(\theta))$.

Spherical coordinates:

$$\int_{\Omega} f(x)\,dx = \int_{\tilde{\Omega}} f\big(r\sin(\varphi)\cos(\theta), r\sin(\varphi)\sin(\theta), r\cos(\varphi)\big) r^2 \sin(\varphi)\,dr\,d\theta\,d\varphi,$$

where $(r, \theta, \varphi) \to x$ is a one-to-one mapping of $\tilde{\Omega}$ onto Ω given by $x = (r\sin(\varphi)\cos(\theta), r\sin(\varphi)\sin(\theta), r\cos(\varphi))$.

75.15 Surface integrals

If $S = s(\Omega)$ is a surface in $\mathbb{R}^3$ parameterized by the mapping $s : \Omega \to \mathbb{R}^3$, where Ω is a domain in $\mathbb{R}^2$, and $u : S \to \mathbb{R}$ is a real-valued function defined on S, then

$$\int_S u\,ds = \int_{\Omega} u(s(y)) \|s'_{,1}(y) \times s'_{,2}(y)\|\,dy,$$

where $s'_{,i} = (\frac{\partial s_1}{\partial y_i}, \frac{\partial s_2}{\partial y_i}, \frac{\partial s_3}{\partial y_i})$.

75.16 Green's and Gauss' formulas

If Ω is a domain in $\mathbb{R}^3$ with boundary Γ with outward unit normal $n = (n_1, n_2, n_3)$, and $u : \Omega \to \mathbb{R}^3$ and $v, w : \Omega \to \mathbb{R}$, then

$$\int_{\Omega} \frac{\partial v}{\partial x_i}\,dx = \int_{\Gamma} v\,n_i\,ds, \quad i = 1, 2, 3.$$

$$\int_{\Omega} \frac{\partial v}{\partial x_i} w\,dx = \int_{\Gamma} vw\,n_i\,ds - \int_{\Omega} v \frac{\partial w}{\partial x_i}\,dx, \quad i = 1, 2, 3.$$

$$\int_{\Omega} \nabla \cdot u\,dx = \int_{\Gamma} u \cdot n\,ds, \quad \text{Gauss}$$

$$\int_{\Omega} \nabla \times u\,dx = \int_{\Gamma} n \times u\,ds,$$

$$\int_{\Omega} \nabla v \cdot \nabla w\,dx = \int_{\Gamma} v \partial_n w\,ds - \int_{\Omega} v \Delta w\,dx,$$

$$\int_{\Omega} v \Delta w\,dx - \int_{\Omega} \Delta v\,w\,dx = \int_{\Gamma} v\,\partial_n w\,ds - \int_{\Gamma} \partial_n v\,w\,ds.$$

75.17 Stokes' theorem

If S is a surface in $\mathbb{R}^3$ bounded by a closed curve Γ, n is a unit normal to S, Γ is oriented in a clockwise direction following the positive direction of the normal n, and $u : \mathbb{R}^3 \to \mathbb{R}^3$ is differentiable, then

$$\int_S (\nabla \times u) \cdot n \, ds = \int_\Gamma u \cdot ds.$$

76
Piecewise linears in $\mathbb{R}^2$ and $\mathbb{R}^3$

...usually he sat in a comfortable attitude, looking down, slightly stooped, with hands folded above his lap. He spoke quite freely, very clearly, simply and plainly: but when he wanted to emphasise a new viewpoint ... then he lifted his head, turned to one of those sitting next to him, and gazed at him with his beautiful, penetrating blue eyes during the emphatic speech. ... If he proceeded from an explanation of principles to the development of mathematical formulas, then he got up, and in a stately very upright posture he wrote on a blackboard beside him in his peculiarly beautiful handwriting: he always succeeded through economy and deliberate arrangement in making do with a rather small space. For numerical examples, on whose careful completion he placed special value, he brought along the requisite data on little slips of paper. (Dedekind about Gauss)

76.1 Introduction

In this chapter, we prepare for the application of FEM to partial differential equations by discussing approximation of functions by piecewise linear functions in in $\mathbb{R}^2$ and $\mathbb{R}^3$. We consider three main topics: (i) the construction of a mesh, or *triangulation*, for a domain in $\mathbb{R}^2$ or $\mathbb{R}^3$, (ii) the construction piecewise linear functions on a triangulation, and (iii) estimation of interpolation errors.

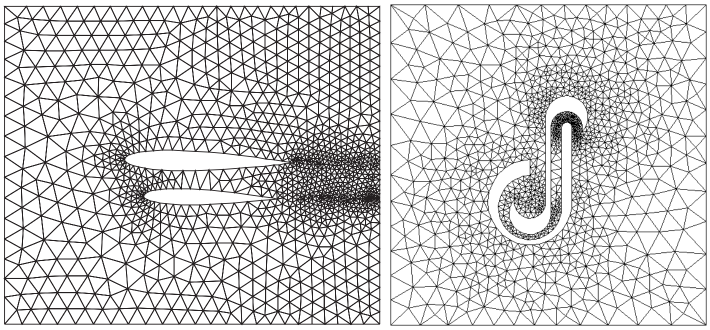


FIGURE 76.1. The mesh on the left was used in a computation of the flow of air around two airfoils. The mesh on the right was used to discretize a piece of metal punched with a fancy character. In both cases, the meshes are adapted to allow accurate computation, taking into account both the behavior of the solution and the shape of the domain.

76.2 Triangulation of a domain in $\mathbb{R}^2$

We start by considering a two-dimensional domain Ω with a polygonal boundary Γ. A *triangulation* $\mathcal{T}_h = \{K\}$ is a sub-division of Ω into a non-overlapping set of triangles, or *elements*, K constructed so that no vertex of one triangle lies on the edge of another triangle, see Fig. 76.2. We use $\mathcal{N}_h = \{N\}$ to denote the set of *nodes* N or corners of the triangles, usually numbered $N_1, N_2, ..., N_M$, where M is the total number of nodes. A triangulation is specified by a list of the coordinates of the nodes, together with a list containing the numbers of the nodes of each triangle. We may also list the set of triangle sides or *edges* $\mathcal{S}_h = \{S\}$, with each edge S specified by the node numbers of its two end-points, and a list of the nodes and edges on the boundary Γ.

We measure the size of a triangle $K \in \mathcal{T}_h$, by the length h_K of its largest side, which is called the *diameter* of the triangle. The *mesh function* $h(x)$ associated to a triangulation $\mathcal{T}_h$ is the piecewise constant function defined so $h(x) = h_K$ for $x \in K$ for each $K \in \mathcal{T}_h$. We measure the degree of *isotropy* of an element $K \in \mathcal{T}_h$ by its smallest angle τ_K. If $\tau_K \approx \pi/3$ then K is almost isosceles, while if τ_K is small then K is thin, see Fig. 76.3. We use the smallest angle among the triangles in $\mathcal{T}_h$, i.e.

$$\tau = \min_{K \in \mathcal{T}_h} \tau_K$$

as a measure of the degree of anistropy of the triangulation $\mathcal{T}_h$. We shall see below that certain interpolation errors related to approximation with piecewise linear functions on a given triangulation get larger as τ tends to zero, corresponding to allowing the triangles to very thin.

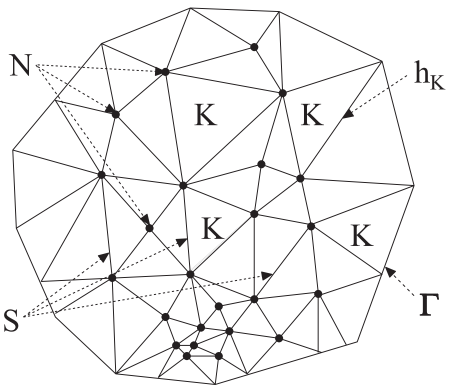


FIGURE 76.2. A triangulation of a domain Ω.

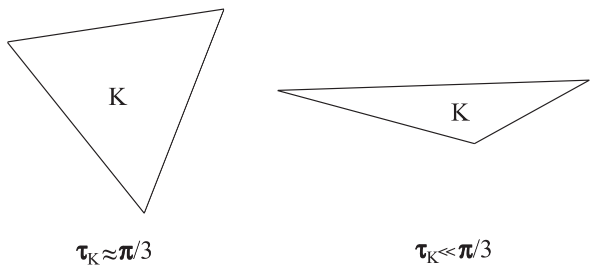


FIGURE 76.3. Measuring the isotropy of a triangle.

The basic problem of *mesh generation* is to generate a triangulation of a given domain with mesh size given by a prescribed mesh function $h(x)$. This problem arises in each step of an adaptive algorithm, where a new mesh function is computed from an approximate solution on a given mesh, and a new mesh is constructed with mesh size given by the new mesh function. The process is then repeated until a stopping criterion is satisfied. The new mesh may be constructed from scratch or by modification of the previous mesh including local refinement or coarsening.

In the *advancing front* strategy a mesh with given mesh size is constructed beginning at some point (often on the boundary) by successively adding one triangle after another, each with a mesh size determined by the mesh function. The curve dividing the domain into a part already triangulated and the remaining part is called the *front*. The front sweeps through the domain during the triangulation process. An alternative is to use a *h-refinement* strategy, where a mesh with a specified local mesh size is constructed by successively dividing elements of an initial coarse triangulation with the elements referred to as *parents*, into smaller elements, called the *children*. We illustrate the refinement and advancing front strategies in

Fig. 76.4. It is often useful to combine the two strategies using the advanc-

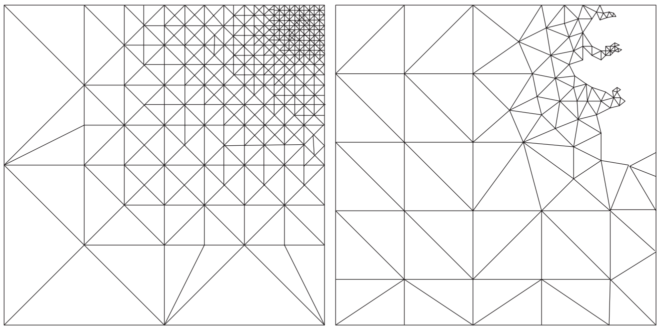


FIGURE 76.4. The mesh on the left is being constructed by successive h refinement starting from the coarse parent mesh drawn with thick lines. The mesh on the right is being constructed by an advancing front strategy. In both cases, high resolution is required near the upper right-hand corner.

ing front strategy to construct an initial mesh that represents the geometry of the domain with adequate accuracy, and use adaptive h-refinement.

There are various strategies for performing the division in an h-refinement aimed at limiting the degree of anisotropy of the elements. After the refinements are completed, the resulting mesh is fixed up by the addition of edges aimed at avoiding nodes that are located in the middle of element sides. This causes a mild "spreading" of the adapted region. We illustrate one technique for h-refinement in Fig. 76.5. In general, refining a mesh tends

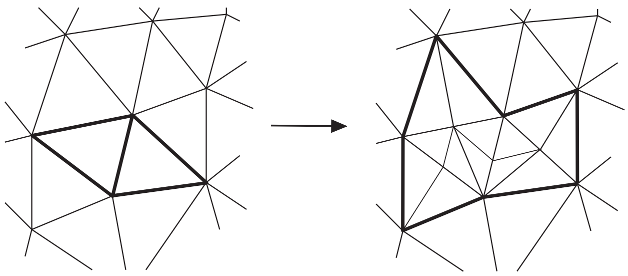


FIGURE 76.5. On the left, two elements in the mesh have been marked for refinement. The refinement uses the Rivara algorithm in which an element is divided into two pieces by inserting a side connecting the node opposite the longest side to the midpoint of the longest side. Additional sides are added to avoid having a node of one element on the side of another element. The refinement is shown in the mesh on the right along with the boundary of all the elements that had to be refined in addition to those originally marked for refinement.

to introduce elements with small angles, as can be seen in Fig. 76.5 and it is an interesting problem to construct algorithms for mesh refinement that avoid this tendency in situations where the degree of anisotropy has to be limited. On the other hand, in certain circumstances, it is important to use "stretched" meshes that have regions of thin elements aligned together to give a high degree of refinement in one direction. In these cases, we also introduce mesh functions that give the local stretching, or degree of anisotropy, and the orientation of the elements. We discuss the construction and use of such meshes in the advanced companion volume.

76.3 Mesh generation in $\mathbb{R}^3$

Mesh generation in three dimensions is analogous to that in two dimensions with the triangles being replaced by *tetrahedra*. In practice, the geometric constraints involved become more complicated and the number of elements also increases drastically. We show some examples in Fig. **??** and Fig. 76.7.

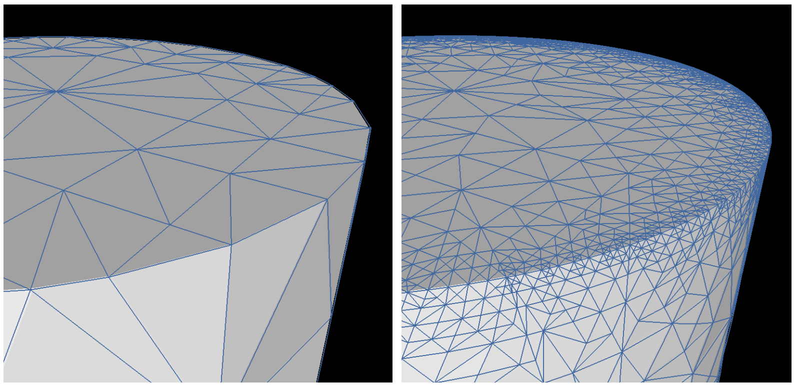


FIGURE 76.6. Initial and refined mesh of cylinder.

76.4 Piecewise linear functions

Let $\mathcal{T}_h = \{K\}$ be a triangulation of a two-dimensional domain Ω with piecewise polynomial boundary Γ, let $\mathcal{N}_h = \{N\}$ denote the nodes of $\mathcal{T}_h$ and introduce the corresponding the finite dimensional vector space V_h consisting of the continuous piecewise linear functions on $\mathcal{T}_h$. In other words,

$$V_h = \{v : v \text{ is continuous on } \Omega, v|_K \in \mathcal{P}(K) \text{ for } K \in \mathcal{T}_h\},$$

where $\mathcal{P}(K)$ denotes the set of linear functions on K, i.e., the set of functions $v(x) = v(x_1, x_2)$ of the form $v(x) = c_0 + c_1 x_1 + c_2 x_2$ for some constants

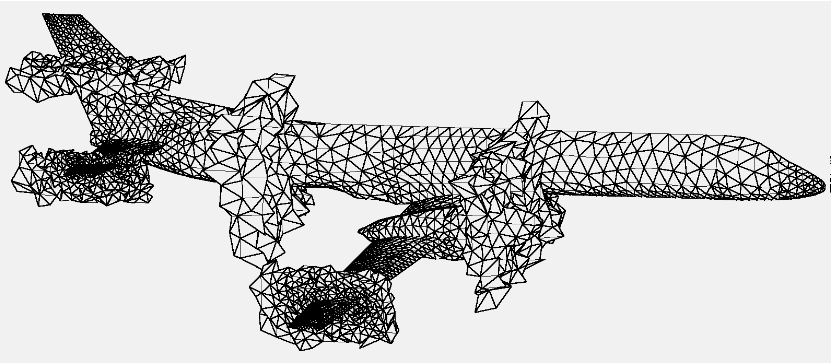


FIGURE 76.7. The surface mesh on the body, and parts of a tetrahedral mesh around a Saab 2000

c_i. We can describe a function $v(x)$ in V_h by the nodal values $v(N)$ with $N \in \mathcal{N}_h$ because of two facts. The first is that a linear function is uniquely determined by its values at three points, as long as they don't lie on a straight line. To prove this claim, let $K \in \mathcal{T}_h$ have vertices $a^i = (a_1^i, a_2^i)$,

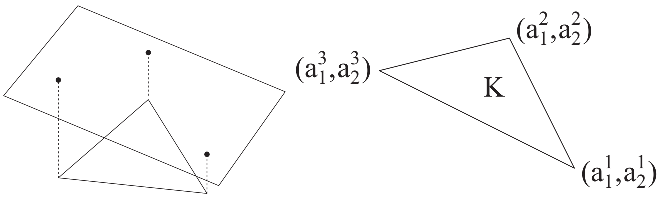


FIGURE 76.8. On the left, we show that the three nodal values on a triangle determine a linear function. On the right, we show the notation used to describe the nodes of a typical triangle.

$i = 1, 2, 3$, see Fig. 76.8. We want to show that $v \in \mathcal{P}(K)$ is determined uniquely by $\{v(a^1), v(a^2), v(a^3)\} = \{v_1, v_2, v_3\}$. A linear function v can be written $v(x_1, x_2) = c_0 + c_1 x_1 + c_2 x_2$ for some constants c_0, c_1, c_2. Substituting the nodal values of v into this expression yields a linear system of equations:

$$\begin{pmatrix} 1 & a_1^1 & a_2^1 \\ 1 & a_1^2 & a_2^2 \\ 1 & a_1^3 & a_1^3 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}.$$

The determinant of the coefficient matrix is equal to the determinant of the following matrix resulting from subtracting the first row from the second

and third row:
$$\begin{pmatrix} 1 & a_1^1 & a_2^1 \\ 0 & a_1^2 - a_1^1 & a_2^2 - a_2^1 \\ 0 & a_1^3 - a_1^1 & a_2^2 - a_2^1 \end{pmatrix},$$
which is equal to the twice the area of the triangle K (up to the sign). The determinant of the coefficient matrix is thus non-zero, and we conclude that the system of equations(76.4) has a unique solution. We conclude that at linear function is uniquely specified by its values at three non-colinear points.

The second fact is that if a function is linear in each of two neighboring triangles and its nodal values on the two common nodes of the triangles are equal, then the function is continuous across the common edge. To see this,

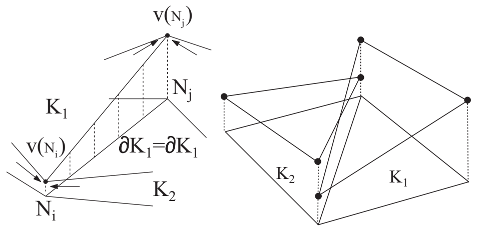


FIGURE 76.9. On the left, we show that a function that is piecewise linear on triangles reduces to a linear function of one variable on triangle edges. On the right, we plot a function that is piecewise linear on triangles whose values at the common nodes on two neighboring triangles do not agree.

let K_1 and K_2 be adjoining triangles with common boundary $\partial K_1 = \partial K_2$; see the figure on the left in Fig. 76.9. Parametrizing v along this boundary, we see that v is a linear function of one variable there. Such functions are determined uniquely by the value at two points, and therefore since the values of v on K_1 and K_2 at the common nodes agree, the values of v on the common boundary between K_1 and K_2 agree, and v is indeed continuous across the boundary.

To construct a set of basis functions for V_h, we begin by describing a set of *element basis functions* for triangles. Once again, assuming that a triangle K has nodes at $\{a^1, a^2, a^3\}$, the element nodal basis is the set of functions $\lambda_i \in \mathcal{P}(K)$, $i = 1, 2, 3$, such that
$$\lambda_i(a^j) = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$
We show these functions in Fig. 76.10.

We construct the *global* basis functions for V_h by piecing together the element basis functions on neighboring elements using the continuity requirement, i.e. by matching element basis functions on neighboring trian-

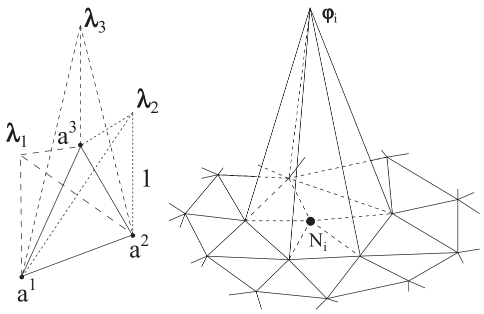


FIGURE 76.10. On the left, we show the three element nodal basis functions for the linear functions on K. On the right, we show a typical global basis "tent" function.

gles that have the same nodal values on the common edge. The resulting set of basis functions $\{\varphi_j\}_{j=1}^M$, where N_1, N_2,..., N_M is an enumeration of the nodes $N \in \mathcal{N}_h$, is called the set of *tent* functions. The tent functions can also be defined by specifying that $\varphi_j \in V_h$ satisfy

$$\varphi_j(N_i) = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$$

for $i, j = 1, ..., M$. We illustrate a typical tent function in Fig. 76.10. We see in particular that the *support* of φ_i is the set of triangles that share the common node N_i.

The tent functions are a nodal basis for V_h because if $v \in V_h$ then

$$v(x) = \sum_{i=1}^{M} v(N_i)\varphi_i(x).$$

76.5 Max-norm error estimates for piecewise linear interpolation

In this section we prove the basic pointwise maximum norm error estimate for linear interpolation on a triangle, which states that the interpolation error depends on the second order partial derivatives of the function being interpolated, i.e. on the "curvature" of the function, the mesh size and the shape of the triangle. Analogous results hold for other norms. The results also extend directly to more than two space dimensions.

Let K be a triangle with vertices $a^i, i = 1, 2, 3$. Given a continuous function v defined on K, let the linear interpolant $\pi_K v \in \mathcal{P}(K)$ be defined by

$$\pi_K v(a^i) = v(a^i), \quad i = 1, 2, 3.$$

We illustrate this in Fig. 76.11.

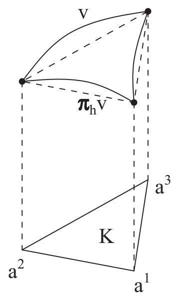


FIGURE 76.11. The nodal interpolant $\pi_K v$ of v.(change from π_h to π_K)

Theorem 76.1 *If v has continuous second derivatives, then*

$$\|v - \pi_K v\|_{L_\infty(K)} \leq \frac{1}{2} h_K^2 \|D^2 v\|_{L_\infty(K)}, \tag{76.1}$$

$$\|\nabla(v - \pi_K v)\|_{L_\infty(K)} \leq \frac{3}{\sin(\tau_K)} h_K \|D^2 v\|_{L_\infty(K)}, \tag{76.2}$$

where h_K is the largest side of K, τ_K is the smallest angle of K, and

$$D^2 v = \left(\sum_{i,j=1}^{2} \left(\frac{\partial^2 v}{\partial x_i \partial x_j}\right)^2 \right)^{1/2}.$$

If ∇v is continuous, then

$$\|v - \pi_K v\|_{L_\infty(K)} \leq h_K \|Dv\|_{L_\infty(K)}, \tag{76.3}$$

Note that the gradient estimate depends on the reciprocal of the sine of the smallest angle of K, and therefore this error bound deteriorates as the the triangle gets thinner.

The proof follows the same general outline as the proofs of corresponding results in the Chapter Piecewise linear approximation. Let λ_i, $i = 1, 2, 3$, be the element basis functions for $\mathcal{P}(K)$ defined by $\lambda_i(a^j) = 1$ if $i = j$, and $\lambda_i(a^j) = 0$ otherwise. A function $w \in \mathcal{P}(K)$ has the representation

$$w(x) = \sum_{i=1}^{3} w(a^i)\lambda_i(x) \quad \text{for } x \in K,$$

and thus

$$\pi_K v(x) = \sum_{i=1}^{3} v(a^i)\lambda_i(x) \quad \text{for } x \in K, \tag{76.4}$$

since $\pi_K v(a^i) = v(a^i)$. We shall derive representation formulas for the interpolation errors $v - \pi_K v$ and $\nabla(v - \pi_K v)$, using a Taylor expansion at $x \in K$:

$$v(y) = v(x) + \nabla v(x) \cdot (y - x) + R(x, y),$$

where

$$R(x, y) = \frac{1}{2} \sum_{i,j=1}^{2} \frac{\partial^2 v}{\partial x_i \partial x_j}(\xi)(y_i - x_i)(y_j - x_j),$$

is the remainder term of order 2 and ξ is a point on the line segment between x and y. In particular choosing $y = a^i = (a_1^i, a_2^i)$, we have

$$v(a^i) = v(x) + \nabla v(x) \cdot (a^i - x) + R_i(x), \tag{76.5}$$

where $R_i(x) = R(x, a^i)$. Inserting (76.5) into (76.4) gives for $x \in K$

$$\pi_K v(x) = v(x) \sum_{i=1}^{3} \lambda_i(x) + \nabla v(x) \cdot \sum_{i=1}^{3} (a^i - x)\lambda_i(x) + \sum_{i=1}^{3} R_i(x)\lambda_i(x). \tag{76.6}$$

We shall use the following identities that hold for $j, k = 1, 2$, and $x \in K$,

$$\sum_{i=1}^{3} \lambda_i(x) = 1, \quad \sum_{i=1}^{3} (a_j^i - x_j)\lambda_i(x) = 0, \tag{76.7}$$

$$\sum_{i=1}^{3} \frac{\partial}{\partial x_k}\lambda_i(x) = 0, \quad \sum_{i=1}^{3} (a_j^i - x_j)\frac{\partial \lambda_i}{\partial x_k} = \delta_{jk}, \tag{76.8}$$

where $\delta_{jk} = 1$ if $j = k$ and $\delta_{jk} = 0$ otherwise. The first of the identities in (76.7) follows by choosing $v(x) = 1$ in (76.6), and the second follows by choosing $v(x) = d_1 x_1 + d_2 x_2$ with $d_i \in \mathbb{R}$. Finally, (76.8) follows by differentiating (76.7).

Using (76.7), we obtain the following representation of the interpolation error,

$$v(x) - \pi_K v(x) = -\sum_{i=1}^{3} R_i(x)\lambda_i(x).$$

Since $|a^i - x| \le h_K$, we can estimate the remainder term $R_i(x)$ as

$$|R_i(x)| \le \frac{1}{2} h_K^2 \|D^2 v\|_{L_\infty(K)}, \quad i = 1, 2, 3.$$

where we used Cauchy's inequality twice to estimate an expression of the form $\sum_{ij} x_i c_{ij} x_j = \sum_i x_i \sum_j c_{ij} x_j$.

Now, using the fact that $0 \le \lambda_i(x) \le 1$ if $x \in K$, for $i = 1, 2, 3$, we obtain

$$|v(x) - \pi_K v(x)| \le \max_i |R_i(x)| \sum_{i=1}^{3} \lambda_i(x) \le \frac{1}{2} h_K^2 \|D^2 v\|_{L_\infty(K)} \quad \text{for } x \in K,$$

which proves (76.1).

To prove (76.2), we differentiate (76.4) with respect to $x_k, k = 1, 2$ to get

$$\nabla(\pi_K v)(x) = \sum_{i=1}^{3} v(a^i)\nabla\lambda_i(x),$$

which together with (76.5) and (76.8) gives the following error representation:

$$\nabla(v - \pi_K v)(x) = -\sum_{i=1}^{3} R_i(x)\nabla\lambda_i(x) \quad \text{for } x \in K.$$

We now note that

$$\max_{x\in K} |\nabla\lambda_i(x)| \leq \frac{2}{h_K \sin(\tau_K)},$$

which follows by an easy estimate of the shortest height (distance from a vertex to the opposite side) of K. We now obtain (76.2) and (76.3) finally follows using tha Mean Value theorem. The proof is now complete.

Let now $\mathcal{T}_h = \{K\}$ be a triangulation of a domain Ω with mesh function $h(x)$, and let π_h denote nodal interpolation into the corresponding space of continuous piecewise linear functions V_h on $\mathcal{T}_h$. The interpolation error estimates of Theorem 76.1 then take the form

$$\|v - \pi_h v\|_{L_\infty(\Omega)} \leq \frac{1}{2}\|h^2 D^2 v\|_{L_\infty(\Omega)}, \tag{76.9}$$

$$\|v - \pi_h v\|_{L_\infty(\Omega)} \leq \|hDv\|_{L_\infty(\Omega)}, \tag{76.10}$$

$$\|\nabla(v - \pi_h v)\|_{L_\infty(\Omega)} \leq \frac{3}{\sin(\tau)}\|hD^2 v\|_{L_\infty(\Omega)}, \tag{76.11}$$

where τ is the smallest of the τ_K. Below we shall use analogs of these estimates with the $L_\infty(\Omega)$ replaced by $L_2(\Omega)$.

76.6 Quadrature in $\mathbb{R}^2$

To compute the stiffness matrix and load vector a FEM, we have to compute integrals of the form $\int_K g(x)\,dx$, where K is a triangle or tetrahedron and g a given function. Sometimes we may evaluate these integrals exactly, but usually this is either impossible or inefficient. Instead we usually evaluate the integrals approximately using quadrature formulas. We briefly present some quadrature formulas for integrals over triangles.

In general, we would like to use quadrature formulas that do not affect the accuracy of the underlying finite element method, which of course requires an estimate of the error due to quadrature. A quadrature formula for an integral over an element K has the form

$$\int_K g(x)\,dx \approx \sum_{i=1}^{q} g(y^i)\omega_i, \tag{76.12}$$

for a specified choice of *nodes* $\{y^i\}$ in K and *weights* $\{\omega_i\}$. We now list some possibilities using the notation a_K^i to denote the vertices of a triangle K, a_K^{ij} to denote the midpoint of the side connecting a_K^i to a_K^j, and a_K^{123} to denote the center of mass of K, and denote by $|K|$ the area of K:

$$\int_K g\,dx \approx g\big(a_K^{123}\big)|K|, \tag{76.13}$$

$$\int_K g(x)\,dx \approx \sum_{j=1}^{3} g(a_K^j)\frac{|K|}{3}, \tag{76.14}$$

$$\int_K g\,dx \approx \sum_{1\le i<j\le 3} g\big(a_K^{ij}\big)\frac{|K|}{3}, \tag{76.15}$$

$$\int_K g\,dx \approx \sum_{j=1}^{3} g\big(a_K^i\big)\frac{|K|}{20} + \sum_{1\le i<j\le 3} g\big(a_K^{ij}\big)\frac{2|K|}{15} + g\big(a_K^{123}\big)\frac{9|K|}{20}. \tag{76.16}$$

We refer to (76.13) as the center of gravity quadrature, to (76.14) as the vertex quadrature, and to (76.15) as the midpoint quadrature. Recall that the accuracy of a quadrature formula is related to the *precision* of the formula. A quadrature formula has precision r if the formula gives the exact value of the integral if the integrand is a polynomial of degree at most $r-1$, but there is some polynomial of degree r such that the formula is not exact. The quadrature error for a quadrature rule of precision r is proportional to h^r, where h is the mesh size. More precisely, the error of a quadrature rule of the form (76.12) satisfies

$$\left|\int_K g\,dx - \sum_{i=1}^{q} g(y^i)\omega_i\right| \le Ch_K^r \sum_{|\alpha|=r}\int_K |D^\alpha g|\,dx,$$

where C is a constant. Vertex and center of gravity quadrature have precision 2, midpoint quadrature has precision 3, while (76.16) has precision 4.

In finite element methods based on continuous piecewise linear functions, we often use nodal or vertex quadrature, often also referred to as *lumped mass* quadrature, because the mass matrix computed this way becomes diagonal.

EXAMPLE 76.1. In Fig. 76.12 and Fig. 76.13 we give two examples, one from electromagnetics and the other from fluid mechanics.

Chapter 76 Problems

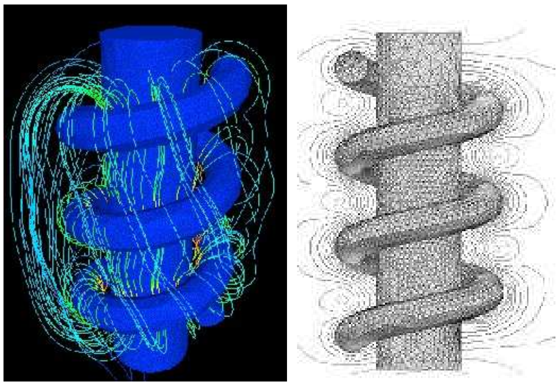


FIGURE 76.12. Magnetic field around coil and mesh

76.1. For a given triangle K, determine the relation between the smallest angle τ_K, the triangle diameter h_K and the diameter ρ_K of the largest inscribed circle.

76.2. Draw the refined mesh that results from sub-dividing the smallest two triangles in the mesh on the right in Fig. 76.5.

76.3. Let K be a tetrahedron with vertices $\{a^i, i = 1, ..., 4\}$. Show that a linear polynomial $v(x) = c_0 + c_1x_1 + c_2x_2 + c_3x_3$ on K is uniquely determined by the nodal values $\{v(a^i), i = 1, ..., 4\}$. Show that the corresponding finite element space V_h consists of continuous functions.

76.4. Prove that the quadrature formulas (76.13),(76.14), (76.15) and (76.16) have the indicated precision.

76.5. Prove that using nodal quadrature to compute a mass matrix for piecewise linears, gives a diagonal mass matrix where a diagonal term is the sum of the terms in the corresponding row in the exactly computed mass matrix. Motivate the term "lumped".

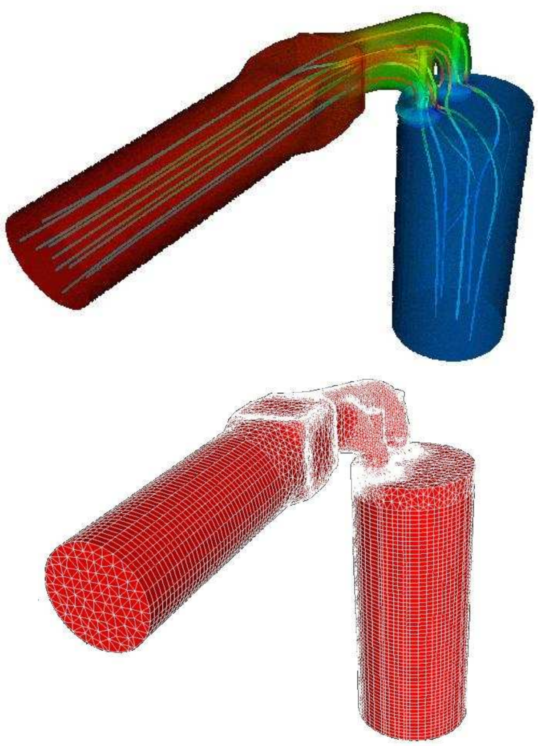


FIGURE 76.13. Flow in diesel engine cylinder inlet and mesh

77

FEM for boundary value problems in $\mathbb{R}^2$ and $\mathbb{R}^3$

...were very confused, skipping suddenly from one idea to another, from one formula to the next, with no attempt to give a connection between them. His presentations were obscure clouds, illuminated from time to time by flashes of pure genius. ... of the thirty who enrolled with me, I was the only one to see it through. (Menabrea about Cauchy 1832)

77.1 Introduction

In this chapter, we extend the cG(1) FEM for reaction-diffusion-convection problems in one space dimension to corresponding boundary value problems in $\mathbb{R}^2$ and $\mathbb{R}^3$ of the form: Find $u : \Omega \rightarrow \mathbb{R}$ such that

$$-\nabla \cdot (a\nabla u) + \nabla \cdot (ub) + cu = f \quad \text{in } \Omega, \tag{77.1}$$

together with boundary conditions of Dirichlet, Neumann or Robin type, where $a(x) > 0$, $b(x)$ and $c(x)$ are given variable coefficients, $f(x)$ is a given right hand side, and Ω is a bounded open domain in $\mathbb{R}^2$ or $\mathbb{R}^3$. Note that the coefficient b is a vector (typically corresponding to a given convection velocity), and that the equation (77.1) can alternatively be written

$$-\nabla \cdot (a\nabla u) + b \cdot \nabla u + \hat{c}u = f \quad \text{in } \Omega, \tag{77.2}$$

with $\hat{c} = c + \nabla \cdot b$. In general, problems of this form cannot be solved analytically and we have to rely on a numerical method such as FEM for computing the solution $u(x)$ for given data.

We consider below the extension to corresponding time dependent problems of the form

$$\dot{u} - \nabla \cdot (a\nabla u) + \nabla \cdot (ub) + cu = f, \tag{77.3}$$

together with initial and boundary value problems, including extensions to systems of such equations, using the material in the Chapters The general initial value problem and Adaptive IVP-solvers.

The most fundamental example of the form (77.1) is Poisson's equation with homogeneous Dirichlet boundary conditions corresponding to setting $a = 1$, $b = 0$ and $c = 0$:

$$\begin{cases} -\Delta u(x) = f(x) & \text{for } x \in \Omega, \\ u(x) = 0 & \text{for } x \in \Gamma, \end{cases} \tag{77.4}$$

where Ω is a bounded domain in $\mathbb{R}^2$ with polygonal boundary Γ. We recall that $\nabla \cdot (\nabla u) = \Delta u$. We shall now present the cG(1) method for (77.4) generalizing cG(1) for the two-point boundary value problem (53.9), and then extend to the general problem (77.1).

77.2 Variational formulation

We let $\mathcal{T}_h = \{K\}$ be a triangulation of Ω with mesh function $h(x)$ and internal nodes $N_1, ...N_M$, and we let V_h be the corresponding finite element space of continuous piecewise linear functions that vanish on the boundary Γ. We first give (77.4) the following preliminary variational formulation:

$$-\int_\Omega \Delta u\, v\, dx = \int_\Omega f\, v\, dx \tag{77.5}$$

for all suitable test functions v, which results from multiplying (77.4) by $v(x)$ and integrating over Ω. We now want to rewrite the left-hand side to move a derivative from Δu onto v. Assuming that the test function v is zero on Γ, Green's formula implies

$$-\int_\Omega \Delta u\, v\, dx = -\int_\Gamma \partial_n uv\, ds + \int_\Omega \nabla u \cdot \nabla v\, dx = \int_\Omega \nabla u \cdot \nabla v\, dx,$$

where ∂_n denotes the outward unit normal derivative on Γ. We find that a solution $u(x)$ of (77.4) satisfies

$$\int_\Omega \nabla u \cdot \nabla v\, dx = \int_\Omega f\, v\, dx, \tag{77.6}$$

for all test functions v with $v = 0$ on Γ.

77.3 The cG(1) FEM

We are thus led to the following formulation of the cG(1) FEM for (77.4): Find $U \in V_h$ such that

$$\int_\Omega \nabla U \cdot \nabla v \, dx = \int_\Omega f \, v \, dx \quad \text{for all } v \in V_h, \tag{77.7}$$

where V_h is the space of continuous piecewise linear functions on a triangulation $\mathcal{T}_h$ of Ω that vanish on the boundary Γ. Using the notation

$$(w, v) = \int_\Omega wv \, dx, \quad (\nabla w, \nabla v) = \int_\Omega \nabla w \cdot \nabla v \, dx,$$

we can write cG(1) in the form: Find $U \in V_h$ such that

$$(\nabla U, \nabla v) = (f, v) \quad \text{for all } v \in V_h. \tag{77.8}$$

We see that that the trial space and test spaces are equal ($= V_h$) and include the homogenous Dirichlet boundary condition. The Galerkin orthogonality is expressed by

$$(\nabla u - \nabla U, \nabla v) = 0 \quad \text{for all } v \in V_h, \tag{77.9}$$

which results upon subtracting (77.8) from (77.6) with $v \in V_h$.

We recall that the nodal basis functions $\{\varphi_i\}_{i=1}^M$ associated with the internal nodes $N_1, ...N_M$ of $\mathcal{T}_h$ is a basis for V_h. Expressing U in terms of this basis,

$$U(x) = \sum_{j=1}^M U(N_j)\varphi_j(x), \tag{77.10}$$

substituting into (77.8), and choosing $v = \varphi_i$ for $i = 1, ..., M$, gives

$$\sum_{j=1}^M (\nabla\varphi_j, \nabla\varphi_i) U(N_j) = (f, \varphi_i), \quad i = 1, ..., M.$$

This is equivalent to the linear system of equations

$$A\xi = b, \tag{77.11}$$

where $\xi = (\xi_j)$ is the vector of internal nodal values $\xi_j = U(N_j)$, $A = (a_{ij})$ is the *stiffness matrix* with elements $a_{ij} = (\nabla\varphi_j, \nabla\varphi_i)$ and $b = (b_i)$ with $b_i = (f, \varphi_i)$ is the *load vector*..

The stiffness matrix A is obviously symmetric and it is also positive-definite since for any $v = \sum_{i=1}^{M} \eta_i \varphi_i$ in V_h,

$$\sum_{i,j=1}^{M} \eta_i a_{ij} \eta_j = \sum_{i,j=1}^{M} \eta_i (\nabla\varphi_i, \nabla\varphi_j)\eta_j$$
$$= \left(\nabla \sum_{i=1}^{M} \eta_i \varphi_i, \nabla \sum_{j}^{M} \eta_j \varphi_j \right) = (\nabla v, \nabla v) > 0,$$

unless $\eta_i = 0$ for all i. This means in particular that (77.11) has a unique solution vector U and thus the cG(1) finite element problem (77.8) has a unique solution $U \in V_h$.

77.3.1 An example: uniform triangulation of a square

We compute the stiffness matrix A and load vector b in (77.11) explicitly on the uniform triangulation of the square $\Omega = [0,1] \times [0,1]$ pictured in Fig. 77.1. We choose an integer $m \geq 1$ and set $h = 1/(m+1)$, then construct the triangles as shown. The diameter of the triangles in $\mathcal{T}_h$ is $\sqrt{2}h$ and there are $M = m^2$ internal nodes. We number the nodes starting from the lower left and moving right, then working up across the rows.

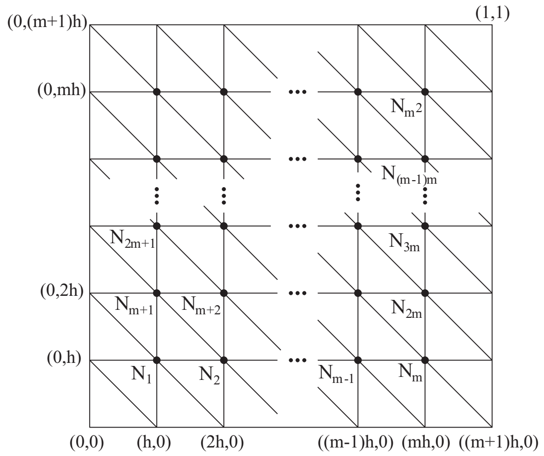


FIGURE 77.1. The standard triangulation of the unit square.

In Fig. 77.2, we show the support of the basis function corresponding to the node N_i along with parts of the basis functions for the neighboring nodes. As in one dimension, the basis functions are "almost" orthogonal in the sense that only basis functions φ_i and φ_j sharing a common triangle

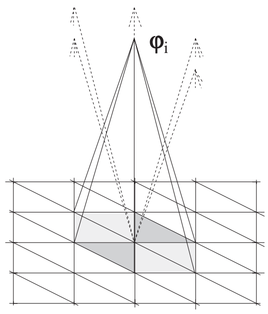


FIGURE 77.2. The support of the basis function φ_i together with parts of the neighboring basis functions.

in their supports yield a non-zero value in $(\nabla\varphi_i, \nabla\varphi_j)$. We show the nodes neighboring N_i in Fig. 77.3. The support of any two neighboring basis

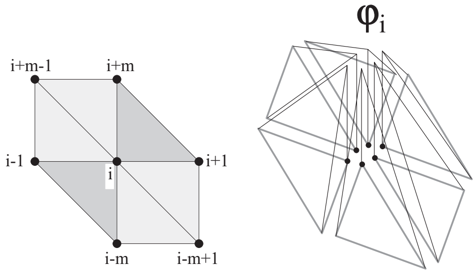


FIGURE 77.3. The indices of the nodes neighboring N_i and an "exploded" view of φ_i.

functions overlap on just two triangles, while a basis function "overlaps itself" on six triangles.

We first compute

$$a_{ii} = (\nabla\varphi_i, \nabla\varphi_i) = \int_\Omega |\nabla\varphi_i|^2\,dx \; = \int\limits_{\text{support of }\varphi_i} |\nabla\varphi_i|^2\,dx,$$

for $i = 1, ..., m^2$. As noted, we only have to consider the integral over the domain pictured in Fig. 77.3, which is written as a sum of integrals over the six triangles making up the domain. Examining φ_i on these triangles, see Fig. 77.3, we see that there are only two different integrals to be computed since φ_i looks the same, except for orientation, on two of the six triangles and similarly the same on the other four triangles. We shade the corresponding triangles in Fig. 77.2. The orientation affects the direction of $\nabla\varphi_i$ of course, but does not affect $|\nabla\varphi_i|^2$.

We compute $(\nabla\varphi_i, \nabla\varphi_i)$ on the triangle shown in Fig. 77.4. In this case,

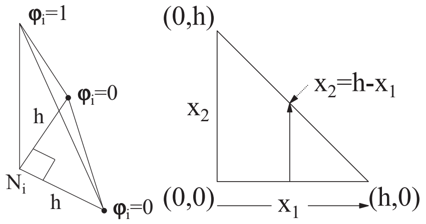


FIGURE 77.4. First case showing φ_i on the left together with the variables used in the reference triangle.

φ_i is one at the node located at the right angle in the triangle and zero at the other two nodes. We change coordinates to compute $(\nabla\varphi_i, \nabla\varphi_i)$ on the *reference triangle* shown in Fig. 77.4. Again, changing to these coordinates does not affect the value of $(\nabla\varphi_i, \nabla\varphi_i)$ since $\nabla\varphi_i$ is constant on the triangle. On the triangle, φ_i can be written $\varphi_i = ax_1 + bx_2 + c$ for some constants a, b, c. Since $\varphi_i(0,0) = 1$, we get $c = 1$. Similarly, we compute a and b to find that $\varphi_i = 1 - x_1/h - x_2/h$ on this triangle. Therefore, $\nabla\varphi_i = \left(-h^{-1}, -h^{-1}\right)$ and the integral is

$$\int_{\triangleright} |\nabla\varphi_i|^2\, dx = \int_0^h \int_0^{h-x_1} \frac{2}{h^2}\, dx_2\, dx_1 = 1.$$

In the second case, φ_i is one at a node located at an acute angle of the triangle and is zero at the other nodes. We illustrate this in Fig. 77.5. We

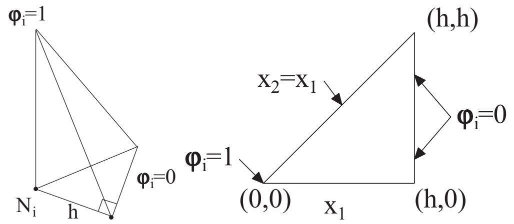


FIGURE 77.5. Second case showing φ_i and the reference triangle.

use the coordinate system shown in Fig. 77.5 to write $\varphi_i = 1 - x_1/h$. When we integrate over the triangle, we get $1/2$.
Summing the contributions from all the triangles gives

$$a_{ii} = (\nabla\varphi_i, \nabla\varphi_i) = 1 + 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 4.$$

Next, we compute $(\nabla\varphi_i, \nabla\varphi_j)$ for indices corresponding to neighboring nodes. For a general node N_i, there are two cases of inner products (see

Fig. 77.2 and Fig. 77.3):

$$a_{i\,i-1} = (\nabla\varphi_i, \nabla\varphi_{i-1}) = (\nabla\varphi_i, \nabla\varphi_{i+1}) = (\nabla\varphi_i, \nabla\varphi_{i-m}) = (\nabla\varphi_i, \nabla\varphi_{i+m}),$$

and

$$a_{i\,i-m+1} = (\nabla\varphi_i, \nabla\varphi_{i-m+1}) = (\nabla\varphi_i, \nabla\varphi_{i+m-1}).$$

The orientation of the triangles in each of the two cases are different, but the inner product of the gradients of the respective basis functions is not affected by the orientation. Note that the the equations corresponding to nodes next to the boundary are special, because the nodal values on the boundary are zero, see Fig. 77.1. For example, the equation corresponding to N_1 only involves N_1, N_2 and N_{m+1}.

For the first case, we next compute $(\nabla\varphi_i, \nabla\varphi_{i+1})$. Plotting the intersection of the respective supports shown in Fig. 77.6, we conclude that there

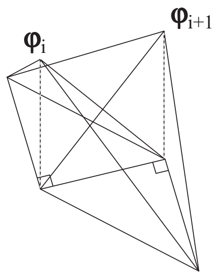


FIGURE 77.6. The overlap of φ_i and φ_{i+1}.

are equal contributions from each of the two triangles in the intersection. We choose one of the triangles and construct a reference triangle as above. Choosing suitable variables, we find that

$$\nabla\varphi_i \cdot \nabla\varphi_{i+1} = \left(-\frac{1}{h}, -\frac{1}{h}\right) \cdot \left(\frac{1}{h}, 0\right) = -\frac{1}{h^2},$$

and integrating over the triangle gives $-1/2$. Similarly, we see that

$$(\nabla\varphi_i, \nabla\varphi_{i-m+1}) = (\nabla\varphi_i, \nabla\varphi_{i+m-1}) = 0.$$

We can now determine the stiffness matrix A using the information above. We start by considering the first row. The first entry is $(\nabla\varphi_1, \nabla\varphi_1) = 4$ since N_1 has no neighbors to the left or below. The next entry is $(\nabla\varphi_1, \nabla\varphi_2) = -1$. The next entry after that is zero, because the supports of φ_1 and φ_3 do not overlap. This is true in fact of all the entries up to and including φ_m. However, $(\nabla\varphi_1, \nabla\varphi_{m+1}) = -1$, since these neighboring basis functions do share two supporting triangles. Finally, all the rest of the entries in that row are zero because the supports of the corresponding basis functions do not overlap. We continue in this fashion working row by row. The result

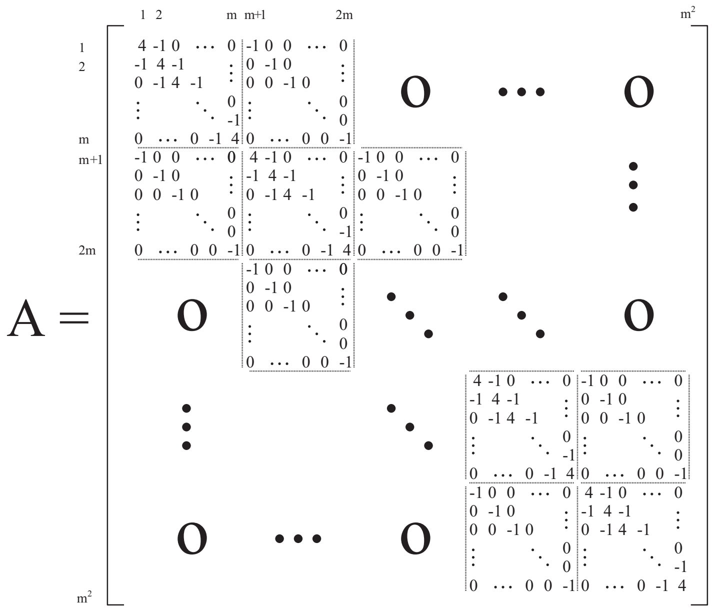


FIGURE 77.7. The stiffness matrix.

is pictured in Fig. 77.7. We see that A has a *block structure* consisting of banded $m \times m$ sub-matrices, most of which consist only of zeros. Note the pattern of entries around corners of the diagonal block matrices; it is a common mistake to program these values incorrectly.

The storage of a sparse matrix and the solution of a sparse system are both affected by the *structure* or *sparsity pattern* of the matrix. The sparsity pattern is affected in turn by the enumeration scheme used to mark the nodes.

There are several algorithms for reordering the coefficients of a sparse matrix to form a matrix with a smaller bandwidth. Reordering the coefficients is equivalent to computing a new basis for the vector space.

The load vector b is computed in the same fashion, separating each integral

$$\int_\Omega f\varphi_i\,dx = \int_{\text{support of } \varphi_i} f(x)\varphi_i(x)\,dx$$

into integrals over the triangles making up the support of φ_i. To compute the elements (f, φ_i) of the load vector, we often use one of the quadrature formulas presented in Chapter 76.

77.4 Basic data structures

To compute the finite element approximation U, we have to compute the coefficients of the stiffness matrix A and load vector b and solve the linear system of equations (77.11). We just computed A and b for a uniform triangulation of the unit square, and we now discuss the case of a general triangulation of a general domain.

We have to compute the non-zero elements $a_{ij} = (\nabla\varphi_j, \nabla\varphi_i)$ of the stiffness matrix A. We know that $a_{ij} = 0$ unless both N_i and N_j are nodes of the same triangle K, because only then the supports of basis functions φ_i and φ_j overlap. The common support corresponding to a non-zero element a_{ij} is equal to the support of φ_i if $j = i$, and is equal to the two triangles with the common edge connecting N_j and N_i if $i \neq j$. In each case a_{ij} is the sum of contributions

$$a_{ij}^K = \int_K \nabla\varphi_j \cdot \nabla\varphi_i \, dx \tag{77.12}$$

over the triangles K in the common support. The process of adding up the contributions a_{ij}^K from the relevant triangles K to get the element a_{ij}, is called *assembling* the stiffness matrix A. Arranging for a given triangle K the numbers a_{ij}^K, where N_i and N_j are nodes of K, into a 3×3 matrix, we obtain the *element stiffness matrix* for the triangle K. We refer to the assembled matrix A as the *global stiffness matrix*. Notice that we use *element* with two different meanings: as an element a_{ij} of the stiffness matrix A, and as a finite element or triangle of the triangulation.

To compute the element stiffness matrix a_{ij}^K for a given triangle K, we need the physical coordinates of the nodes of K. To perform the assembly where we loop over all elements and add the corresponding contributions to the global stiffness matrix, we need the node numbers of each triangle. Similar information is needed to compute the load vector. The required information is arranged in a *data structure*, or data base, containing (i) a list of the coordinates of the nodes numbered in some way, and (ii) a list of the node numbers of each triangle. A list of the numbers of the nodes on the boundary is also needed to handle the boundary conditions This information is typically the output of the mesh generator generating a triangulation of the domain.

77.5 Solving the discrete system

Once we have assembled the stiffness matrix A and computed the load vector b, we have to solve the linear system $AU = b$ to obtain the finite element approximation $U(x)$. We now discuss this topic briefly based on the material presented in Chapter 44. The stiffness matrix resulting from discretizing the Laplacian is symmetric and positive-definite and therefore invertible. These properties also mean that there is a wide choice in the methods used to solve the linear system $AU = b$, which take advantage of the fact that A is sparse.

In the case of the standard uniform discretization of a square, we saw that A is a banded matrix with five non-zero diagonals and bandwidth $m + 1$, where m is the number of nodes on a side. The dimension of A is m^2 and the asymptotic operations count for using Gaussian elimination is $O(m^4) = O(h^{-4})$. Note that even though A has mostly zero diagonals inside the band, fill-in occurs as the elimination is performed, so we may as well treat A as if it has non-zero diagonals throughout the band. Clever rearrangement of A to reduce the amount of fill-in leads to a solution algorithm with an operations count on the order of $O(m^3) = O(h^{-3})$. In contrast, if we treat A as a full matrix, we get an asymptotic operations count of $O(h^{-6})$, which is considerably larger for a large number of elements.

In general, we get a sparse stiffness matrix, though there may not be a band structure. If we want to use a Gaussian elimination method efficiently in general, then it is necessary to first reorder the system to bring the matrix into banded form.

We can also apply both the Jacobi and Gauss-Seidel methods to solve the linear system arising from discretizing the Poisson equation. In the case of the uniform standard discretization of a square for example, the operations count is $O(M)$ per iteration for both methods if we make use of the sparsity of A. Therefore a single step of either method is much cheaper than a direct solve. The question is: How many iterations do we need to compute in order to obtain an accurate solution?

Typically the spectral radius of the iteration matrix of the Jacobi or Gauss-Seidel method is equal to $1 - Ch^2$ with C some moderate positive constant, which means that the convergence rate quickly gets slow as h decreases: to reduce the error a certain factor, we need of the order of $O(h^{-2})$ iterations, and since each iteration takes $O(h^{-2})$ operations, the total number of operations is $O(h^{-4})$, which is the same order as using a banded Gaussian elimination solver.

There has been a lot of activity in developing iterative methods that converge more quickly than Jacobi and Gauss-Seidel. In recent years, very efficient *multi-grid methods* have been developed and are now becoming a standard tool. A multi-grid method is based on a sequence of Gauss-Seidel or Jacobi steps performed on a hierarchy of successively coarser meshes and are optimal in the sense that the solution work is proportional to the total number of unknowns (that is h^{-2} in the model problem).

77.6 An equivalent minimization problem

The variational problem (77.8) is equivalent to the following quadratic *minimization problem*: find $U \in V_h$ such that

$$F(u) \leq F(v) \quad \text{for all } v \in V_h, \tag{77.13}$$

where

$$F(v) = \frac{1}{2}\int_\Omega |\nabla v|^2\,dx - \int_\Omega fv\,dx = \frac{1}{2}(\nabla v, \nabla v) - (f, v).$$

The quantity $F(v)$ may be interpreted as the *total energy* of the function $v(x)$ composed of the *internal energy* $\frac{1}{2}(\nabla v, \nabla v)$ and the *load potential* $-(f,v)$. Thus, the solution U minimizes the total energy $F(v)$ with v varying over V_h.

To see the equivalence of (77.8) and (77.13), we assume first that $U \in V_h$ satisfies (77.8). Let then $v \in V_h$ and write $v = U + (v - U) = U + w$ with $w = v - U \in V_h$. Using $(\nabla U, \nabla w) = (\nabla w, \nabla U)$, we get

$$\begin{aligned}F(v) = F(U+w) = &\frac{1}{2}(\nabla U, \nabla U) + (\nabla U, \nabla w) + \frac{1}{2}(\nabla w, \nabla w) - (f, U) - (f, w)\\ =&F(U) + \frac{1}{2}(\nabla w, \nabla w) \geq F(U),\end{aligned}$$

with equality only if $w = 0$. We conclude that U satisfies (77.13).

Conversely, if U is the solution of (77.13), then we have for all $v \in V_h$

$$g_v(\epsilon) = F(U + \epsilon v) \geq g(0) = F(U) \quad \text{for all } \epsilon \in \mathbb{R},$$

and thus $\epsilon = 0$ is an interior minimum point of $g_v(\epsilon)$ with v fixed, and thus $g_v'(0) = 0$. Computing we get

$$0 = g_v'(0) = (\nabla U, \nabla v) - (f, v)$$

and thus U satisfies (77.8). We sum up in the following theorem:

Theorem 77.1 *The problems* (77.8) *and* (77.13) *are equivalent in the sense that they have the same unique solution.*

77.7 An energy norm a priori error estimate

In this section, we derive a priori and a posteriori estimates of the error $u - U$ in the *energy norm* $\|\nabla(u - U\|$ with

$$\|\nabla v\| = (\int_\Omega |\nabla v|^2\,dx)^{1/2}, \tag{77.14}$$

where u is the exact solution and U a finite element solution of Poisson's equation with homogeneous Dirichlet boundary conditions. The energy norm, which is the L_2 norm of the gradient of a function in this problem, arises naturally in the error analysis of the finite element method because it is closely tied to the variational problem. The gradient of the solution, representing heat flow, electric field, flow velocity, or stress for example, can be a variable of physical interest as much as the solution itself, representing temperature, potential or displacement for example, and in this case, the energy norm is the relevant error measure.

We first prove that the Galerkin finite element approximation is the best approximation of the true solution in V_h with respect to the energy norm.

Theorem 77.2 *Assume that u satisfies the Poisson equation* (77.4) *and U is the Galerkin finite element approximation satisfying* (77.8). *Then*

$$\|\nabla(u-U)\| \leq \|\nabla(u-v)\| \quad \textit{for all } v \in V_h. \tag{77.15}$$

The proof goes as follows: Using the Galerkin orthogonality (77.9) with v replaced by $U - v \in V_h$, we can write

$$\|\nabla e\|^2 = (\nabla e, \nabla(u-U)) = (\nabla e, \nabla(u-U)) + (\nabla e, \nabla(U-v)).$$

Adding the terms involving U on the right, whereby U drops out, and using Cauchy's inequality, we get

$$\|\nabla e\|^2 = (\nabla e, \nabla(u-v)) \leq \|\nabla e\|\, \|\nabla(u-v)\|,$$

which proves the theorem after dividing by $\|\nabla e\|$.

Choosing $v = \pi_h u$ and using an $L_2(\Omega)$ analog of the interpolation estimate (76.11), we get the following quantitative a priori error estimate (with $\|v\| = \|v\|_{L_2(\Omega)}$):

Corollary 77.3 *There exists a constant C_i depending only on the minimal angle τ in $\mathcal{T}_h$, such that*

$$\|\nabla(u-U)\| \leq C_i \|hD^2u\|. \tag{77.16}$$

77.8 An energy norm a posteriori error estimate

We now prove an a posteriori error estimate following the strategy used for the two-point boundary value problem in Chapter 53. A new feature occurring in higher dimensions is the appearance of integrals over the internal edges S in $\mathcal{S}_h$. We start by writing an equation for the error $e = u - U$ using (77.6) and (77.8) to get

$$\begin{aligned}\|\nabla e\|^2 &= (\nabla(u-U), \nabla e) = (\nabla u, \nabla e) - (\nabla U, \nabla e)\\ &= (f, e) - (\nabla U, \nabla e) = (f, e - \pi_h e) - (\nabla U, \nabla(e - \pi_h e)),\end{aligned}$$

where $\pi_h e \in V_h$ is an interpolant of e. We now break up the integrals over Ω into sums of integrals over the triangles K in $\mathcal{T}_h$ and integrate by parts over each triangle in the last term to get

$$\|\nabla e\|^2 = \sum_K \int_K (f + \Delta U)(e - \tilde{\pi}_h e)\,dx - \sum_K \int_{\partial K} \frac{\partial U}{\partial n_K}(e - \pi_h e)\,ds, \quad (77.17)$$

where $\partial U/\partial n_K$ denotes the derivative of U in the outward normal direction n_K of the boundary ∂K of K. In the boundary integral sum in (77.17), each internal edge $S \in \mathcal{S}_h$ occurs twice as a part of each of the boundaries ∂K of the two triangles K that have S as a common side. Of course the outward normals n_K from each of the two triangles K sharing S point in opposite directions. For each side S, we choose one of these normal directions and denote by $\partial_S v$ the derivative of a function v in that direction on S. We note that if $v \in V_h$, then in general $\partial_S v$ is different on the two triangles sharing S; see Fig. 76.9, which indicates the "kink" over S in the graph of v. We can express the sum of the boundary integrals in (77.17) as a sum of integrals over edges of the form

$$\int_S [\partial_S U](e - \pi_h e)\,ds,$$

where $[\partial_S U]$ is the difference, or jump, in the derivative $\partial_S U$ computed from the two triangles sharing S. The jump appears because the outward normal directions of the two triangles sharing S are opposite. We further note that $e - \tilde{\pi}_h e$ is continuous across S, but in general does not vanish on S even if π_h is the nodal interpolant. This is different than the one-dimensional case, where the corresponding sum over nodes does indeed vanish because $e - \pi_h e$ vanishes at the nodes. We may thus rewrite (77.17) as follows with the second sum replaced by a sum over internal edges S:

$$\|\nabla e\|^2 = \sum_K \int_K (f + \Delta U)(e - \pi_h e)\,dx + \sum_{S \in \mathcal{S}_h} \int_S [\partial_S U](e - \pi_h e)\,ds.$$

Next, we return to a sum over element edges ∂K by just distributing each jump equally to the two triangles sharing it, to obtain an *error representation* of the energy norm of the error in terms of the residual error:

$$\|\nabla e\|^2 = \sum_K \int_K (f + \Delta U)(e - \pi_h e)\,dx + \sum_K \frac{1}{2} \int_{\partial K} h_K^{-1} [\partial_S U](e - \pi_h e) h_K\,ds,$$

where we have prepared to estimate the second sum by inserting a factor h_K and compensating. In crude terms, the residual error results from substituting U into the differential equation $-\Delta u - f = 0$, but in reality straightforward substitution is not possible because U is not twice differentiable in Ω. The integral on the right over K is the remainder from substituting U into the differential equation inside each triangle K, while the integral over ∂K arises because $\partial_S U$ in general is different when computed from the two triangles sharing S.

We estimate the first term in the error representation by inserting a factor h, compensating and using the estimate $\|h^{-1}(e-\pi_h e)\| \le C_i \|\nabla e\|$ analogous to (76.11), to obtain

$$
\begin{aligned}
&|\sum_K \int_K h(f+\Delta U)h^{-1}(e-\pi_h e)\,dx| \\
&\qquad\qquad \le \|hR_1(U)\|\|h^{-1}(e-\pi_h e)\| \le C_i \|hR_1(U)\|\|\nabla e\|,
\end{aligned}
$$

where $R_1(U)$ is the function defined on Ω by setting $R_1(U) = |f+\Delta U|$ on each triangle $K \in \mathcal{T}_h$. We estimate the contribution from the jumps on the edges similarly. Formally, the estimate results from replacing $h_K\,ds$ by dx corresponding to replacing the integrals over element boundaries ∂K by integrals over elements K. Dividing by $\|\nabla e\|$, we obtain the following a posteriori error estimate:

Theorem 77.4 *There is an interpolation constant C_i only depending on the minimal angle τ such that the error of the Galerkin finite element approximation U of the solution u of the Poisson equation satisfies*

$$
\|\nabla u - \nabla U\| \le C_i \|hR(U)\|, \tag{77.18}
$$

where $R(U) = R_1(U) + R_2(U)$ with

$$
\begin{aligned}
R_1(U) &= |f+\Delta U| \quad \text{on } K \in \mathcal{T}_h, \\
R_2(U) &= \frac{1}{2}\max_{S\subset\partial K} h_K^{-1}\big|[\partial_S U]\big| \quad \text{on } K \in \mathcal{T}_h.
\end{aligned}
$$

Note that $R_1(U)$ is the contribution to the total residual from the interior of the elements K. In the present case of piecewise linear approximation, $R_1(U) = |f|$. Further, $R_2(U)$ is the contribution to the residual from the jump of the normal derivative of U across edges. In the one dimensional problem considered in Chapter 53, this contribution does not appear because the interpolation error may be chosen to be zero at the node points.

77.9 Adaptive error control

The basic goal of adaptive error control is to find a triangulation $\mathcal{T}_h$ with a least number of nodes such that the corresponding finite element approximation U satisfies

$$
\|\nabla u - \nabla U\| \le \text{TOL}. \tag{77.19}
$$

Using the a posteriori error estimate we are thus led to find a triangulation $\mathcal{T}_h$ with a least number of nodes such that the corresponding finite element approximation U satisfies

$$C_i \|hR(U)\| \leq \text{TOL}. \tag{77.20}$$

This is a nonlinear constrained minimization problem with U depending on $\mathcal{T}_h$. If (77.18) is a reasonably sharp estimate of the error, then a solution of this optimization problem will meet our original goal.

We cannot expect to be able to solve this minimization problem analytically. Instead, a solution has to be sought by an iterative process in which we start with a coarse initial mesh and then successively modify the mesh by seeking to satisfy the stopping criterion (77.20) with a minimal number of elements. More precisely, we follow the following *adaptive algorithm*:

1. Choose an initial triangulation $\mathcal{T}_h^{(0)}$.
2. Given the j^{th} triangulation $\mathcal{T}_{h^{(j)}}$ with mesh function $h^{(j)}$, compute the corresponding finite element approximation $U^{(j)}$.
3. Compute the corresponding residuals $R_1(U^{(j)})$ and $R_2(U^{(j)})$ and check whether or not (77.20) holds. If it does, stop.
4. Find a new triangulation $\mathcal{T}_{h^{(j+1)}}$ with mesh function $h^{(j+1)}$ and with a minimal number of nodes such that $C_i\|h^{(j+1)}R(U^{(j)})\| \leq \text{TOL}$, and then proceed to #2.

The success of this iteration hinges on the mesh modification strategy used to perform step #4. A natural strategy for error control based on the L_2 norm uses the *principle of equidistribution* of the error in which we try to equalize the contribution from each element to the integral defining the L_2 norm. The rationale is that refining an element with large contribution to the error norm gives a large pay-off in terms of error reduction per new degree of freedom.

In other words, the approximation computed on the optimal mesh $\mathcal{T}_h$ in terms of computational work satisfies

$$\|\nabla e\|_{L_2(K)}^2 \approx \frac{\text{TOL}^2}{M} \quad \text{for all } K \in \mathcal{T}_h,$$

where M is the number of elements in $\mathcal{T}_h$. Based on (77.18), we would therefore like to compute the triangulation at step #4 so that

$$C_i^2\big(\big\|h^{(j+1)}R\big(U^{(j+1)}\big)\big\|_{L_2(K)}^2 \approx \frac{\text{TOL}^2}{M^{(j+1)}} \quad \text{for all } K \in \mathcal{T}_{h^{(j+1)}}, \tag{77.21}$$

where $M^{(j+1)}$ is the number of elements in $\mathcal{T}_{h^{(j+1)}}$. However, (77.21) is a nonlinear equation, since we don't know $M^{(j+1)}$ and $U^{(j+1)}$ until we have chosen the triangulation. Hence, we replace (77.21) by

$$C_i^2\big(\big\|h^{(j+1)}R\big(U^{(j)}\big)\big\|_{L_2(K)}^2 \approx \frac{\text{TOL}^2}{M^{(j)}} \quad \text{for all } K \in \mathcal{T}_h{}^{(j+1)}, \tag{77.22}$$

and use this formula to compute the new mesh size $h^{(j+1)}$.

There are several questions that need to be answered about the process described here, including: how much efficiency is lost by replacing (77.19) by (77.20)? In other words, how much bigger is the right-hand side of (77.18) than the left-hand? Does the iterative process #1–#4 converge to a solution of the minimization problem? How should the initial triangulation $\mathcal{T}_{h^{(0)}}$ be chosen and how does this affect the convergence of the adaptive procedure? Is the approximation (77.22) justified? We address these issues in the advanced companion volume.

77.10 An example

We want to compute the the solution

$$u(x) = \frac{a}{\pi}\,\exp\bigl(-a(x_1^2 + x_2^2)\bigr), \quad a = 400,$$

of Poisson's equation $-\Delta u = f$ on the square $(-.5, .5) \times (-.5, .5)$ with $f(x)$ being the following "approximate delta function":

$$f(x) = \frac{4}{\pi}\,a^2\,\bigl(1 - ax_1^2 - ax_2^2\bigr)\,\exp\bigl(-a(x_1^2 + x_2^2)\bigr),$$

We plot f in Fig. 77.8 (note the vertical scale), together with the initial mesh with 224 elements. The adaptive algorithm took 5 steps to achieve

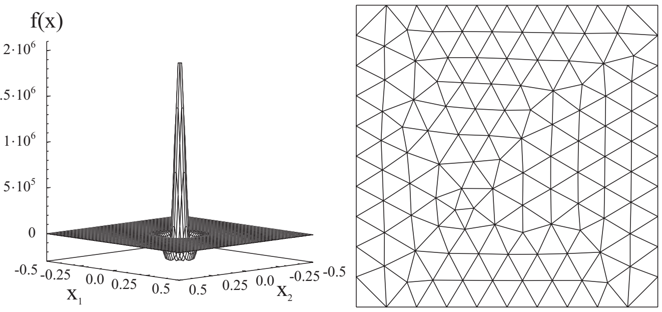


FIGURE 77.8. The approximate delta forcing function f and the initial mesh used for the finite element approximation.

an estimated .5% relative error. We plot the final mesh together with the associated finite element approximation in Fig. 77.9. The algorithm produced meshes with 224, 256, 336, 564, 992, and 3000 elements respectively.

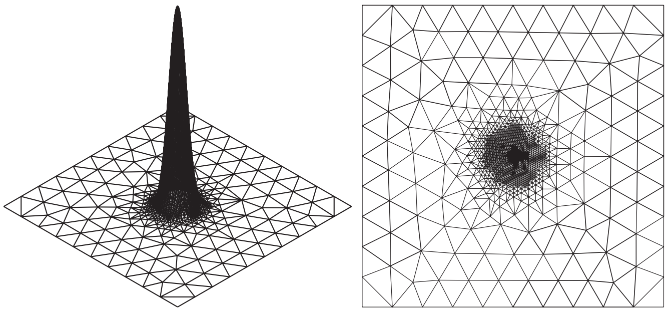


FIGURE 77.9. The finite element approximation with a relative error of .5% and the final mesh used to compute the approximation. The approximation has a maximum height of roughly 5.

77.11 Extension to non-homogeneous Dirichlet boundary conditions

We now consider Poisson's equation with non-homogeneous Dirichlet boundary conditions:

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = g & \text{on } \Gamma, \end{cases} \tag{77.23}$$

where g is the given boundary data.

We compute a finite element approximation on a triangulation $\mathcal{T}_h$, where we now also include the nodes on the boundary, denoting the internal nodes by $\mathcal{N}_h$ as above and the set of nodes on the boundary by $\mathcal{N}_b$. We compute an approximation U of the form

$$U = \sum_{N_j \in \mathcal{N}_b} \xi_j \varphi_j + \sum_{N_j \in \mathcal{N}_h} \xi_j \varphi_j, \tag{77.24}$$

where φ_j denotes the basis function corresponding to node N_j in an enumeration $\{N_j\}$ of all the nodes, and, because of the boundary conditions, $\xi_j = g(N_j)$ for $N_j \in \mathcal{N}_b$. Thus the boundary values of U are given by g on Γ and only the coefficients of U corresponding to the interior nodes remain to be found. To this end, we substitute (77.24) into the variational formulation of (77.23) with the test functions being all the basis functions for the internal nodes, and we then get the following a square system of linear equations for the unknown coefficients of U:

$$\sum_{N_j \in \mathcal{N}_h} \xi_j (\nabla\varphi_j, \nabla\varphi_i) = (f, \varphi_i) - \sum_{N_j \in \mathcal{N}_b} g(N_j)(\nabla\varphi_j, \nabla\varphi_i), \quad N_i \in \mathcal{N}_h.$$

where the terms with known boundary values of U are shifted to the right hand side as data.

77.12 The L-shaped membrane

We present an example that shows the performance of the adaptive algorithm on a problem with a *boundary singularity* with the derivatives of the exact solution being infinite at a corner of the boundary. We consider the Laplace equation in an L-shaped domain that has a non-convex corner at the origin satisfying homogeneous Dirichlet boundary conditions at the sides meeting at the origin and non-homogeneous conditions on the other sides, see Fig. 77.10. We choose the boundary conditions so that the exact solution is $u(r,\theta) = r^{2/3}\sin(2\theta/3)$ in polar coordinates (r,θ) centered at the origin, which has a typical singularity of a corner problem:

$$\frac{\partial u}{\partial r}(r,\theta) = \frac{2}{3}r^{-1/3}\sin(2\theta/3),$$

which tends to infinity as r tends to zero (unless $\theta = 0$ or $\theta = \frac{3\pi}{2}$).

We use the knowledge of the exact solution to evaluate the performance of the adaptive algorithm.

We compute using an adaptive FEM-solver with energy norm control based on (77.18) to achieve an error tolerance of TOL$=$ .005 using h refinement mesh modification. In Fig. 77.10, we show the initial mesh $\mathcal{T}_{h^{(0)}}$ with 112 nodes and 182 elements. In Fig. 77.11, we show the level curves

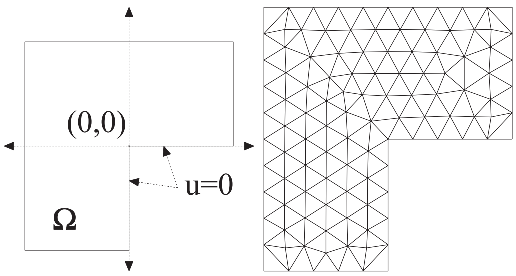


FIGURE 77.10. The L-shaped domain and the initial mesh.

of the solution and the final mesh with 295 nodes and 538 elements that achieves the desired error bound. The interpolation constant was set to $C_i = 1/8$. The quotient between the estimated and true error on the final mesh was 1.5.

Since the exact solution is known in this example, we can also use the a priori error estimate to determine a mesh that gives the desired accuracy. We do this by combining the a priori error estimate (77.16) and the principle of equidistribution of error to determine $h(r)$ so that $C_i\|hD^2u\| =$ TOL while keeping h as large as possible (and keeping the number of elements at a minimum). Since $D^2u(r) \approx r^{-4/3}$, as long as $h \le r$, that is up to the

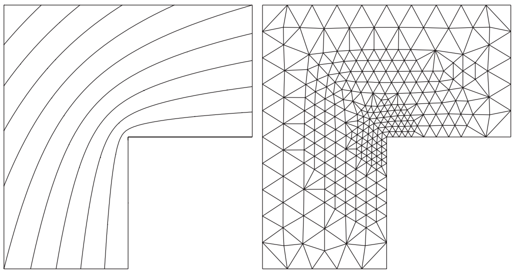


FIGURE 77.11. Level curves of the solution and final adapted mesh on the L-shaped domain.

elements touching the corner, we determine that

$$\left(hr^{-4/3}\right)^2 h^2 \approx \frac{\mathrm{TOL}^2}{M} \qquad \text{or} \quad h^2 = \mathrm{TOL}\, M^{-1/2} r^{4/3},$$

where M is the number of elements and h^2 measures the element area. To compute M from this relation, we note that $M \approx \int_\Omega h^{-2}\, dx$, since the number of elements per unit area is $O(h^{-2})$, which gives

$$M \approx M^{1/2} \mathrm{TOL}^{-1} \int_\Omega r^{-4/3}\, dx.$$

Since the integral is convergent (prove this), it follows that $M \propto \mathrm{TOL}^{-2}$, which implies that $h(r) \propto r^{1/3}\,\mathrm{TOL}$. Note that the total number of unknowns, up to a constant, is the same as that required for a smooth solution without a singularity, namely TOL^{-2}. This depends on the very local nature of the singularity in the present case. In general, of course solutions with singularies may require a much larger number of elements than smooth solutions do.

77.13 Robin and Neumann boundary conditions

Next, we consider Poisson's equation with homogeneous Dirichlet conditions on part Γ_1 of the boundary and non-homogeneous Robin conditions on the remaining part of the boundary Γ_2:

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ u = 0 & \text{on } \Gamma_1, \\ \partial_n u + \kappa u = g & \text{on } \Gamma_2, \end{cases} \tag{77.25}$$

where $\kappa \geq 0$ is a given coefficient, and f and g are given data. Setting $\kappa = 0$ gives the Neumann condition $\partial_n u + \kappa u = g$. To find a variational formulation, we multiply the Poisson equation by a test function v satisfying the homogenous Dirichlet boundary condition, integrate over Ω, and use Green's formula to move derivatives from u to v:

$$\begin{aligned}(f,v) &= -\int_\Omega \Delta u\, v\, dx = \int_\Omega \nabla u \cdot \nabla v\, dx - \int_\Gamma \partial_n u v\, ds\\ &= \int_\Omega \nabla u \cdot \nabla v\, dx + \int_{\Gamma_2} \kappa u v\, ds - \int_{\Gamma_2} g v\, ds,\end{aligned}$$

where we use the boundary conditions to rewrite the boundary integral. We are thus led to the following cG(1) FEM based on a space V_h of continuous piecewise linear functions vanishing on Γ_1: find $U \in V_h$ such that

$$(\nabla U, \nabla v) + \int_{\Gamma_2} \kappa U v\, ds = (f,v) + \int_{\Gamma_2} g v\, ds \quad \text{for all } v \in V_h. \tag{77.26}$$

We recall that boundary conditions like the Dirichlet condition that are enforced explicitely in the choice of the space V_h are called *essential boundary conditions*. Boundary conditions like the Robin condition that are implicitly contained in the weak formulation are called *natural boundary conditions*. (To remember that we must assume essential conditions: there are two "ss" in assume and essential.)

Note that the stiffness matrix and load vector related to (77.26) contain contributions from both integrals over Ω and Γ_2 related to the basis functions corresponding to the nodes on the boundary Γ_2.

To illustrate, we compute the solution of Laplace's equation with a combination of Dirichlet, Neumann and Robin boundary conditions on the domain shown in Fig. 77.12 using an adaptive FEM-solver. We show the boundary conditions in the illustration. The problem models e.g. stationary heat flow around a hot water pipe in the ground. We show the mesh

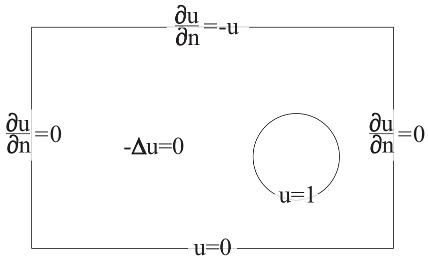


FIGURE 77.12. A problem with Robin boundary conditions.

used to compute the approximation so that the error in the L_2 norm is smaller than .0013 together with a contour plot of the approximation in Fig. 77.13. We notice that the level curves are parallel to a boundary with a homogeneous Dirichlet condition, and orthogonal to a boundary with a homogeneous Neumann condition.

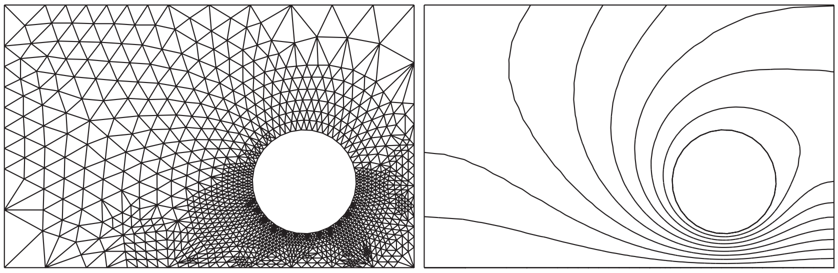


FIGURE 77.13. The adaptive mesh and contour lines of the approximate solution of the problem shown in Fig. 77.12 computed with error tolerance .0013.

77.14 Stationary convection-diffusion-reaction problems

We now consider the extension to a convection-diffusion-reaction problem of the form

$$\begin{aligned} -\nabla\cdot(a\nabla u)+\nabla\cdot(ub)+cu&=f \quad \text{in } \Omega,\\ a\frac{\partial u}{\partial n}+\kappa u&=g \quad \text{on } \Gamma. \end{aligned} \tag{77.27}$$

with Robin boundary conditions, where f and g are given data, and $a>0$, b, c and $\kappa\geq 0$ are given coefficients, and Ω is a given domain in $\mathbb{R}^2$ with boundary Γ. The term cu models absorption if $c\geq 0$ and production $c<0$.

Let V_h be the space of continuous piecewise linear functions on a triangulation of Ω with no restriction on the nodal values on the boundary. The cG(1) FEM for (77.27) takes the form: Find $U\in V_h$ such that

$$\int_\Omega a\nabla U\cdot\nabla v\,dx+\int_\Omega \nabla\cdot(ub)v\,dx+\int_\Omega cuv\,dx+\int_\Gamma \kappa uv\,ds=\int_\Omega fv\,dx+\int_\Gamma gv\,ds, \tag{77.28}$$

for all $v\in V_h$. Note that the extension to include the terms $\nabla\cdot(ub)$ and cu is very natural and that the corresponding terms in the variational formulation are obtained by multiplying by the test function v without any partial integration. For the term $-\nabla\cdot(a\nabla u)$ we note that multiplication by $v(x)$ and integration over Ω gives using the Divergence theorem

$$-\int_\Omega \nabla\cdot(a\nabla u)v\,dx=\int_\Omega a\nabla U\cdot\nabla v\,dx-\int_\Gamma a\frac{\partial u}{\partial n}v\,ds$$

and the variational formulation results from replacing $-a\frac{\partial u}{\partial n}v\,ds$ by $ku-g$ using the Robin boundary condition.

The matrix equation corresponding to (77.28) has a banded and sparse stiffness matrix, but the symmetry is lost if $b\neq 0$, as is evident from the presence of the non-symmetric term $\int_\Omega \nabla\cdot(ub)v\,dx$. The non-symmetry of the convection term eliminates the best approximation property of FEM, but FEM still may give good results. If $|b| > \frac{a}{h}$, where $h(x)$ is the mesh size, and the exact solution is non-smooth with rapid variation, then the FEM-solution may exhibit spurious oscillations. In such cases the cG(1)-method (77.28) will have to be modified, see the CDE-book. If $c < 0$ then solutions may be non-unique corresponding to non-zero solutions (eigen-functions) of the homogeneous problem $-\nabla\cdot(a\nabla u) + \nabla\cdot(ub) + cu = 0$.

77.15 Time-dependent convection-diffusion-reaction problems

We now consider the time-dependent analog of (77.27), that is the problem

$$\begin{aligned} \dot{u} - \nabla\cdot(a\nabla u) + \nabla\cdot(ub) + cu = f \quad &\text{in } \Omega\times(0,T],\\ a\frac{\partial u}{\partial n} + \kappa u = g \quad &\text{on } \Gamma\times(0,T],\\ u(\cdot) = u^0 \quad &\text{in } \Omega, \end{aligned} \tag{77.29}$$

where $[0,T]$ is a given time interval, and u^0 a given initial value. For the time discretization we may use e.g. dG(0) or cG(1) on a subdivision $0 = t_0 < t_1 < \cdots < t_N = T$ into time intervals $I_n = (t_{n-1}, t_n]$ with time steps $k_n = t_n - t_{n-1}$. Using dG(0) we seek $U^n \in V_h$ for $n = 1, ..., N$, such that for $n = 1, ..., N$,

$$\begin{aligned} &\int_\Omega U^n v\,dx + \int_{\Omega\times I_n} a\nabla U^n\cdot\nabla v\,dx\,dt\\ &+ \int_{\Omega\times I_n} \nabla\cdot(U^n b)v\,dx\,dt + \int_{\Omega\times I_n} cU^n v\,dx\,dt + \int_{\Gamma\times I_n} \kappa U^n v\,ds\,dt\\ &= \int_\Omega U^{n-1} v\,dx + \int_{\Omega\times I_n} fv\,dx\,dt + \int_{\Gamma\times I_n} gv\,ds\,dt, \end{aligned} \tag{77.30}$$

for all $v \in V_h$, where $U^0 = u^0$. The corresponding discrete system for U^n takes the form

$$M\xi^n + k_n A_n \xi^n = M\xi^{n-1} + k_n b^n$$

where the vector ξ^n contains the nodal values of $U^n \in V_h$, M is the mass matrix related to V_h, A_n is the relevant stiffness matrix connected to the convection-diffusion-reaction terms, and b^n the relevant load vector.

77.16 The wave equation

We now consider the extension to the wave equation with homogeneous Dirichlet boundary conditions:

$$\begin{aligned} \ddot{u} - \Delta u = f \quad & \text{in } \Omega \times (0,T], \\ u = 0 \quad & \text{on } \Gamma \times (0,T], \\ u = u^0, \quad \dot{u} = \dot{u}^0 \quad & \text{in } \Omega, \end{aligned} \tag{77.31}$$

where u^0 andf $\dot{u}^0$ are given initial conditions. As above we let V_h be the set of piecewise linears functions on a triangulation of Ω satisfying the homogeneous Dirichlet boundary conditions, and we let $0 = t_0 < t_1 < \cdots < t_N = T$ be a subdivision of $[0,T]$ into time intervals $I_n = (t_{n-1}, t_n]$ with time steps $k_n = t_n - t_{n-1}$. We appply cG(1) in space and cG(1) in time and seek a discrete solution U in the space of functions W_h spanned by the functions

$$v(x,t) = \sum_{n=0}^{N} \sum_{j=1}^{M} \eta_j^n \varphi_j(x) \psi_n(t),$$

where $\{\varphi_j(x)\}_{j=1}^M$ is a basis for V_h, and $\{\psi_n(t)\}_{n=0}^N$ is a basis for the space of continuous piecewise linear functions on the subdivision $0 = t_0 < t_1 < \cdots < t_N = T$. The corresponding discrete system takes the following explicit form if mass lumping is used in space as well as time and the time step is constant $k_n = k$:

$$\xi^{n+1} = 2\xi^n - \xi^{n-1} + k^2 A \xi^n \quad \text{for } n = 1, ..., N-1,$$

with appropriate starting values ξ^0 and ξ^1 computed from the initial conditions, and A the relevant stiffness matrix related to the Laplacian.

77.17 Examples

We present some examples of systems of non-linear reaction-diffusion-convection equations (77.27) arising in physics, chemistry and biology of the form

$$\dot{u}_i - \nabla \cdot (a_i \nabla u_i) + \nabla \cdot (u_i b_i) + c_i u_i = f_i(u_1, ..., u_d) \quad \text{in } \Omega \times I,\ i = 1, ..., d, \tag{77.32}$$

where the $a_i > 0$, b_i and c_i are given coefficients, and the $f_i : \mathbb{R}^d \to \mathbb{R}$. These systems may be solved numerically by a direct extension of the cG(1) method in space and time presented above. We will return in detail to this issue below. In all the examples, a is a positive constant.

EXAMPLE 77.1. *The bistable equation for ferro-magnetism*

$$\dot{u} - a\Delta u = u - u^3. \tag{77.33}$$

EXAMPLE 77.2. *Superconductivity of fluids*

$$\begin{aligned}\dot{u}_1 - a\Delta u_1 &= (1-|u|^2)u_1,\\ \dot{u}_2 - a\Delta u_2 &= (1-|u|^2)u_2.\end{aligned} \tag{77.34}$$

EXAMPLE 77.3. *Flame propagation*

$$\begin{aligned}\dot{u}_1 - a\Delta u_1 &= -u_1 e^{-\alpha_1/u_2},\\ \dot{u}_2 - a\Delta u_2 &= \alpha_2 u_1 e^{-\alpha_1/u_2},\end{aligned} \tag{77.35}$$

where $\alpha_1, \alpha_2 > 0$ are constants.

EXAMPLE 77.4. *Interaction of two species*

$$\begin{aligned}\dot{u}_1 - a\Delta u_1 &= u_1 M(u_1,u_2),\\ \dot{u}_2 - a\Delta u_2 &= u_2 N(u_1,u_2),\end{aligned} \tag{77.36}$$

where $M(u_1,u_2)$ and $N(u_1,u_2)$ are given functions describing various situations such as (i) predator-prey ($M_{u_2} < 0$, $N_{u_1} > 0$) (ii) competing species ($M_{u_2} < 0$, $N_{u_1} < 0$) and (iii) symbiosis ($M_{u_2} > 0$, $N_{u_1} > 0$).

EXAMPLE 77.5. *Morphogenesis of patterns (zebra)*

$$\begin{aligned}\dot{u}_1 - a\Delta u_1 &= -u_1 u_2^2 + \alpha_1(1-u_1),\\ \dot{u}_2 - a\Delta u_2 &= u_1 u_2^2 - (\alpha_1+\alpha_2)u_2.\end{aligned} \tag{77.37}$$

EXAMPLE 77.6. *Belousov-Zhabotinski reaction in chemical kinetics*

$$\begin{aligned}\dot{u}_1 - a\Delta u_1 &= \alpha_1(u_2 - u_1u_2 + u_1 - \alpha_2 u_2^2),\\ \dot{u}_2 - a\Delta u_2 &= \alpha_1^{-1}(\alpha_3 u_3 - u_2 - u_1 u_2).\\ \dot{u}_3 - a\Delta u_3 &= \alpha_4(u_1 - u_3),\end{aligned} \tag{77.38}$$

where $\alpha \approx 10^2$, $\alpha_2 \approx 10^{-2}$, $\alpha_3 \approx 1$, $\alpha_4 \approx 10^{-1}$.

Chapter 77 Problems

77.1. Compute the coefficients of the mass matrix M on the standard triangulation of the square of mesh size h. Hint: it is possible to use quadrature based on the midpoints of the sides of the triangle because this is exact for quadratic functions. The diagonal terms are $h^2/2$ and the off-diagonal terms are all equal to $h^2/12$. The sum of the elements in a row is equal to h^2.

77.2. Compute the stiffness matrix for cG(1) for the problem $-\Delta u = 1$ in $\Omega = (0,1) \times (0,1)$ with $u = 0$ on the side with $x_2 = 0$ and $\partial_n u + u = 1$ on the other three sides of Ω using the standard triangulation. Note the contribution to the stiffness matrix from the nodes on the boundary.

77.3. Describe the sparsity pattern of the stiffness matrices A for the Poisson equation with homogeneous Dirichlet data on the unit square corresponding to the continuous piecewise linear finite element method on the standard triangulation using the three numbering schemes pictured in Fig. 77.14.

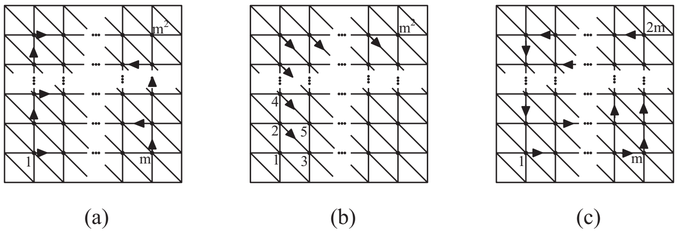


FIGURE 77.14. Three node numbering schemes for the standard triangulation of the unit square.

77.4. Compute the load vector for $f(x) = x_1 + x_2^2$ on the standard triangulation of the unit square using exact integration and the lumped mass (trapezoidal rule) quadrature.

77.5. Write a code to solve $A\xi = b$ using both the Jacobi and Gauss-Seidel iteration methods, making use of the sparsity of A in storage and operations. Compare the convergence rate of the two methods using the result from a direct solver as a reference value.

77.6. Write a code to solve the system $A\xi = b$ with A a band matrix.

77.7. Compute the stiffness matrix for the Poisson equation with homogeneous Dirichlet boundary conditions for (a) the *union jack* triangulation of a square shown in Fig. 77.15 and (b) the triangulation of triangular domain shown in Fig. 77.15.

77.8. Compute the discrete equations for the finite element approximation for $-\Delta u = 1$ on $\Omega = (0,1) \times (0,1)$ with boundary conditions $u = 0$ for $x_1 = 0$, $u = x_1$ for $x_2 = 0$, $u = 1$ for $x_1 = 1$ and $u = x_1$ for $x_2 = 1$ using the standard triangulation (Fig. 77.1).

77.9. (a) Show that the element stiffness matrix (77.12) for the linear polynomials on a triangle K with vertices at $(0,0)$, $(h,0)$, and $(0,h)$ numbered 1, 2 and 3, is given by

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

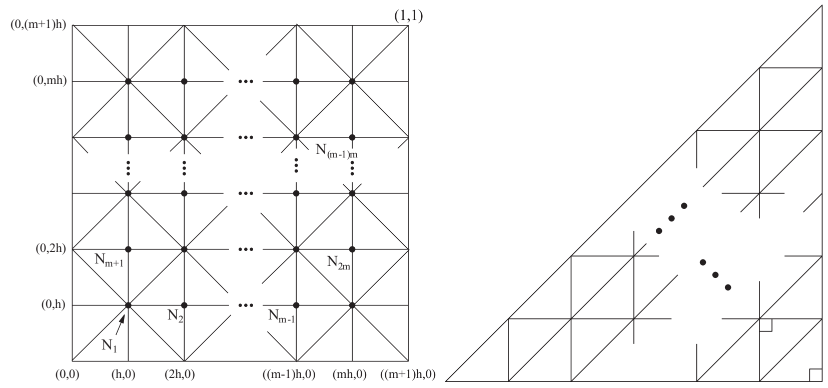


FIGURE 77.15. The “union jack” triangulation of the unit square and a uniform triangulation of a right triangle.

(b) Use this result to verify the formula computed for the stiffness matrix A for the continuous piecewise linear finite element method for the Poisson equation with homogeneous boundary conditions on the unit square using the standard triangulation. (c) Compute the element stiffness matrix for a triangle K with nodes $\{a^i\}$.

77.10. Compute the asymptotic operations count for the direct solution of the system $A\xi = b$ using the three A computed in Problem 77.3.

77.11. Apply the finite element method with piecewise linear approximation to the Poisson equation in three dimensions with a variety of boundary conditions. Compute the stiffness matrix and load vector in some simple cases.

77.12. Derive a priori error bound in the energy norm for cG(1) FEM for Poisson's equation with Robin boundary conditions. Generalize to problems of the form $-\nabla \cdot (a\nabla u) + cu = f$, where $a(x) > 0$ and $c \geq 0$.

77.13. Derive a posteriori error bound in the energy norm for cG(1) FEM for Poisson's equation with Robin boundary conditions. Generalize to problems of the form $-\nabla \cdot (a\nabla u) + cu = f$, where $a(x) > 0$ and $c \geq 0$.

77.14. Implement adaptive energy norm error control for cG(1) for Poisson's equation based on an a posteriori error estimate.

77.15. Find an exact solution of the L-shaped membrane problem with the Dirichlet condition replaced by a Neumann condition on one of the sides meeting at $\frac{3\pi}{2}$ corner. What is the nature of the singularity?

77.16. Let $\omega(x)$ be a positive weight function defined on the domain $\Omega \subset \mathbb{R}^2$. Assume that the mesh function $h(x)$ minimizes the integral $\int_\Omega h^2(x)\omega(x)\,dx$ under the constraint $\int_\Omega h^{-2}(x)\,dx = N$, where N is a given positive integer. Prove that $h^4(x)\omega(x)$ must be constant. Interpret the result as equidistribution in the context of error control. Hint: argue that $h^4(x)\omega(x)$ is the gain adding one more node.

78
Inverse problems

> I never guess. It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.
>
> When you have eliminated the impossible, whatever remains, however improbable, must be the truth. (Sherlock Holmes in the *The Sign of Four*, 1888)

78.1 Introduction

We have above in our study of Poisson's equation studied "forward" problems of the form: Given the function $f : \Omega \to \mathbb{R}$, find a function $u : \Omega \to \mathbb{R}$ such that

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ \partial_n u + \kappa u = 0 & \text{on } \Gamma, \end{cases} \tag{78.1}$$

where $\kappa \geq 0$ is a a given constant. A corresponding "inverse" problem would be to assume knowledge of $u(x)$ and seek the corresponding function $f(x)$ so that (78.19) is satisfied! If we know $u(x)$ in the entire region Ω, this is a problem of differentiation: just compute $\Delta u(x)$ from $u(x)$! We have studied this problem in Chapter *The derivative*, and we recall that this problem is a bit delicate and that we have to balance the step length h in a difference approximation of Δ to the precision in the given data $u(x)$.

Suppose now that we know $u(x)$ only on the boundary Γ. Can we then determine $f(x)$ in Ω? This type of problem connects to a wealth of important applications of the following form: Suppose we can measure something on the *boundary* of an object. Can we then say something about what is *inside* the object? For example, suppose the object is a human body, and that we can measure something on the boundary (or outside) the body. Can we than get some information on what is inside the body? Or, suppose we can accumulate data on the surface of the Earth, can we then say something about what is in the interior of the Earth, such as the presence of layers of oil. These are all examples of *inverse problems*.

The nature of an inverse problem is to be "ill-posed" in the sense that solutions may be non-unique and/or that small changes in the data may cause large changes in the solution. To single out a unique solution which is not too sensitive to little errors in data, we may seek to "regularize" the inverse problem e.g. by smoothing of the data and/or restricting the size of the solution. Differentiation is such an ill-posed problem where we may need to smooth or regularize a given function before attempting to compute its derivative.

A typical forward problem is "well-posed" in the sense that small changes in data cause small changes in the solution. A basic example of a well-posed problem is integration corresponding to solving a differential equation. The corresponding inverse problem is differentiation which is ill-posed as we just noted. Solving a differential equation does not always correspond to a well-posed problem: in Chapter *Lorenz and the Essence of Chaos* we met a simple differential equation with solutions being highly sensitive to changes in data.

EXAMPLE 78.1. An *electrocardiogram* ECG produces a curve reflecting the electrical activity of the heart from measurements of electric potentials on the chest, and the curve gives a specialist information on abnormal activities of the heart such as abnormal heart rythm (arrythmias). Similary, an *electroencephalogram* EEG gives information on the electrical activity of the brain from measurements of electric potentials on the scalp. These techniques are however too imprecise for many diagnoses, and more recently techniques of *electrocardiographic imaging* have been developed, which build on solving inverse problems for Poisson-like equations. The geometry of the individual patient is then obtained from computer tomography, and a picture of the electrical activity inside the body is obtained from measurements of electric potentials on the boundary (e.g. the chest or scalp) by solving an inverse problem for a Poisson-like equation (using the finite element method). Electrocardiographic imaging may give more accurate information on e.g. abnormal cardiac or brain activity than ECG and EEG, and is now a part of practice in advanced neurological and radiological departments. Further development of in particular the computational process (adaptivity, geometric modeling) is needed to increase the accuracy.

EXAMPLE 78.2. Another inverse problem of importance to mankind occurs in inverse seismic prospectation: Explosions on the surface of the Earth are set off and the reflections of the induced waves in the mantle of the Earth are recorded on the surface, and from this information one tries to determine subsurface structures such as layers of oil-bearing rock. To solve this reconstruction problem one uses computational methods based on solving the wave equation involving a wave speed coefficient characteristic of different materials, and through optimization one tries to find the local wave speed coefficient which gives best least squares fit to measured data on the surface of the Earth, and which then gives infomation on the unknown subsurface layering.

78.2 An inverse source problem for one-dimensional convection

We start considering the simplest boundary value problem:

$$u'(x) = f(x), \text{ for } x \in (0,1],\ u(0) = 0, \tag{78.2}$$

modeling convection with $u : [0,1] \to \mathbb{R}$ representing a concentration and $f : [0,1] \to \mathbb{R}$ a source. We seek to determine or *reconstruct* the function $f : [0,1] \to \mathbb{R}$ from the boundary value *observation* $u(1)$ of the corresponding solution $u(x)$ of (78.2). It is clear that we cannot hope to determine $f(x)$ for all $x \in (0,1)$ from this observation alone. This is because there are many functions $f(x)$ such that the corresponding function $u(x)$ satisfies (78.2) and $u(1) = 0$. To see this it is sufficient to choose a non-zero function $u(x)$ on $[0,1]$ satisfying $u(0) = u(1) = 0$ and define $f(x) = -u'(x)$. Evidently, the reconstruction is undetermined up to such functions.

The indeterminancy of the reconstruction $f(x)$ reflects the *ill-posed* nature of the inverse problem; even if the measurements of the boundary value $u(1)$ is very precise, the corresponding source $f(x)$ is not well defined. We thus need some *extra condition* to single out a (hopefully) unique source $f(x)$. We may do this in many ways and depending on the extra condition imposed, we may get different reconstructions $f(x)$. We now indicate one possibility, where we reconstruct under the extra condition that $f(x)$ is *as small as possible* in a least squares sense, which is a common technique of regularization. We then reformulate the inverse problem as the following *least squares optimization problem*: Find the function $f : [0,1] \to \mathbb{R}$ which minimizes the total cost

$$J)f) = (u(1) - \bar{u}(1))^2 + \mu \int_0^1 f(x)^2\, dx,$$

where $u(x)$ solves (78.2), $\bar{u}(1)$ is the observed boundary value at $x = 1$, and $\mu > 0$ is a constant. We may view this as a constrol problem where the objective is to find the *control* $f:[0,1]\to\mathbb{R}$ which mimimizes the total cost $J(f)$, where the μ-term measures the cost of the control f and the first term the cost of a boundary value misfit $u(1)-\bar{u}(1)$.

We can phrase this problem as finding the function $u(x)$ with $u(0)=0$ which minimizes

$$(u(1)-\bar{u}(1))^2+\mu\int_0^1 (u')^2\,dx.$$

Recalling Chapter *FEM for two-point boundary value problems*, we understand that the solution $u(x)$ satisfies $u''(x)=0$ in $(0,1)$, $u(0)=0$ and $u(1)-\bar{u}(1)+\mu u'(1)=0$. We conclude that $u(x)=\frac{1}{1+\mu}\bar{u}(1)x$, and thus the reconstructed source $f(x)=u'(x)$ takes on a constant value and is given by

$$f(x)=\frac{1}{1+\mu}\bar{u}(1)\quad\text{for } x\in[0,1].$$

Evidently, we are led to choose the regularization parameter μ small; the smaller μ is the more accurately we will fit the boundary value observation $\bar{u}(1)$. Since the reconstructed function $f(x)$ is constant, we have effectively only one constant to determine and we may expect to be able determine this single value from the single observation $\bar{u}(1)$.

We can also rephrase the optimization problem as follows introducing the integral operator B defined on functions on $[0,1]$ by $Bf(x)=\int_0^x f(y)\,dy$ for $x\in[0,1]$: Find the function $f:[0,1]\to\mathbb{R}$ which minimizes

$$J(f)=(Bf(1)-\bar{u}(1))^2+\mu\int_0^1 f(x)^2\,dx.$$

The optimality condition obtained by setting $\frac{d}{d\epsilon}J(f+\epsilon g)=0$ for $\epsilon=0$, where $g:[0,1]\to\mathbb{R}$ is an arbitrary function, takes the form:

$$(Bf(1)-\bar{u}(1))Bg(1)+\mu\int_0^1 f(x)g(x)\,dx=0 \tag{78.3}$$

for all functions $g:[0,1]\to\mathbb{R}$. We shall now rewrite this condition by introducing the *adjoint operator* $B^\top$ defined on $\mathbb{R}$ as follows: for a given real number w we let $B^\top w$ be the function on $[0,1]$ satisfying $(B^\top w)'(x)=0$ for $x\in(0,1)$ and $B^\top w(1)=w$, that is, $B^\top w$ is the constant function on $[0,1]$ taking the value w. We can then rewrite (78.4) in the form

$$\int_0^1 B^\top(Bf(1)-\bar{u}(1))(x)g(x)\,dx+\mu\int_0^1 f(x)g(x)\,dx=0 \tag{78.4}$$

for all functions $g:[0,1]\to\mathbb{R}$, because by partial integration

$$\int_0^1 B^\top(Bf(1)-\bar{u}(1))(x)(Bg(x))'\,dx=(Bf(1)-\bar{u}(1))Bg(1),$$

$(Bg(x))' = g(x)$ and $B^\top w(1) = w$. We conclude that

$$\int_0^1 (B^\top Bf(1) + \mu f) g\, dx = \int_0^1 B^\top \bar{u}(1) g\, dx$$

for all functions $g : [0,1] \to \mathbb{R}$, and therefore

$$B^\top Bf(1) + \mu f = B^\top \bar{u}(1) \quad \text{on } (0,1), \tag{78.5}$$

or with I the identity operator:

$$(B^\top B + \mu I) f = B^\top \bar{u}(1). \tag{78.6}$$

We conclude that $f(x)$ is constant on $[0,1]$ and takes the value

$$f(x) = \frac{1}{1+\mu} \bar{u}(1) \quad \text{for } x \in [0,1],$$

which is the same result as already derived. We note the form (78.6) of the optimality condition (78.6) with the operator $B^\top B + \mu I$ appearing. We shall meet the same equation below with different solution operator B and adjoints $B^\top$.

78.3 An inverse source problem for one-dimensional diffusion

We continue with the boundary value problem

$$-u'' = f \quad \text{in } (0,1), \quad u'(0) = 0,\, u'(1) + u(1) = 0, \tag{78.7}$$

where we seek to determine the source $f(x)$ in $(0,1)$ by observing the boundary values $u(0)$ and $u(1)$ of the corresponding solution $u(x)$ of (78.7). Again it is clear that we cannot hope to determine $f(x)$ for all $x \in (0,1)$ from these two observations alone, because there are many functions $f(x)$ such that the corresponding function $u(x)$ satisfies (78.7) and $u(0) = u(1) = 0$. To see this it is sufficient to choose a non-zero function $u(x)$ on $[0,1]$ satisfying $u(0) = u'(0) = u(1) = u'(1) = 0$ and set $f(x) = -u''(x)$.

As above we seek to reconstruct under the extra condition that $f(x)$ is as small as possible and we therefore reformulate the inverse problem as the following least squares optimization problem: Find $f(x)$ in $(0,1)$ such that

$$J(f) = (u(0) - \bar{u}(0))^2 + (u(1) - \bar{u}(1))^2 + \mu \int_0^1 f^2(x)\, dx$$

is as small as possible, where $\mu > 0$ is a positive constant acting as a regularization, $\bar{u}(0)$ and $\bar{u}(1)$ are the boundary observations, and of course

$u(x)$ solves (78.7). We thus seek $f(x)$ so that in a least squares sense we fit the boundary observations as well as the smallness of $f(x)$ as well as possible.

To state the optimality equations, we introduce the solution operator B corresponding to (78.7), that is, for a given function $f:[0,1]\to\mathbb{R}$ we let $Bf(x)$ be the function on $[0,1]$ satisfying

$$\int_0^1 (Bf)'v'\,dx + Bf(1)v(1) = \int_0^1 fv\,dx, \tag{78.8}$$

for all functions $v(x)$ on $[0,1]$. This follows from Chapter *FEM for two-point boundary value problems*. Setting $\frac{d}{d\epsilon}J(f+\epsilon g)=0$ for $\epsilon=0$, we obtain the optimality condition in the form

$$(Bf(0)-\bar{u}(0),Bg(0)) + (Bf(1)-\bar{u}(1),Bg(1)) + \mu\int_0^1 f(x)g(x)\,dx = 0 \tag{78.9}$$

for all functions $g(x)$ on $[0,1]$. Next we introduce the adjoint operator $B^\top$ defined as follows: given the values $w(0)$ and $w(1)$, we let $B^\top w$ be the function on $[0,1]$ which satisfies

$$\int_0^1 (B^\top w)'v'\,dx + B^\top w(1)v(1) = w(0)v(0) + w(1)v(1) \tag{78.10}$$

for all $v(x)$. We see that $(B^\top w)''=0$ and $-(B^\top w)'(0)=w(0)$, $B^\top w(1) + (B^\top w)'(1) = w(1)$. In other words, $B^\top w$ is a linear function determined by the two boundary conditions. In particular, if $w(0)=0$, then $B^\top w = w(1)$ is a constant. Now, setting $v=Bg$ in (78.10), we get

$$\begin{aligned} w(0)Bg(0)+w(1)Bg(1) &= \int_0^1 (B^\top w)'(Bg)'\,dx + B^\top w(1)Bg(1)\\ &= \int_0^1 B^\top wg\,dx, \end{aligned}$$

where we used (78.8) with f replaced by g and v replaced by $B^\top w$, and thus we can write the optimality condition (78.9) in the same form as above:

$$(B^\top B+\mu I)f = B^\top\bar{u}. \tag{78.11}$$

From this equation we can uniquely solve for the function $f(x)$, which will be a linear function defined by two constants, because $f=\frac{1}{\mu}B^\top(Bf-\bar{u})$. For example if $\bar{u}(0)=0$ and $\bar{u}(1)=1$, then we get choosing μ small, $f(x)\approx 10\bar{u}(1)x - 8\bar{u}_1$ with corresponding solution $u(x)\approx -3x^3+4x^2$.

We now comment on the nature of the optimality equation (78.11). The operator B maps a space of sources, say F, into a space of observations, say O, and the adjoint operator $B^\top$ maps O into F. We may think of the dimension of F as large, and that of O as smaller. For the discussion we may assume that the dimension of F is n and the dimension of O is m and thus B corresponds to an $m \times n$ matrix and $B^\top$ to an $n \times m$ matrix with $m << n$. This will be the setting with computational approximations of the solution operators B anbd $B^\top$. In particular, the columns of B must be severely linearly independent since there are many more columns tahn rows, and thus the $n \times n$ matrix $B^\top B$ must be singular with many non-zero n-vectors f satisfying $B^\top B f = 0$. On the other hand, the matrix $B^\top B + \mu I$ with $\mu > 0$ is nonsingular, because if $(B^\top B + \mu I)f = 0$, then scalar multiplication by the n-vector $f^\top$, we obtain $\|Bf\|^2 + \mu\|f|^2 = 0$ and thus $f = 0$. The non-zero solutions f to $B^\top B f = 0$ are eigenvectors corresponding to a zero eigenvalue, and by changing to the regularized operator $B^\top B + \mu I$ we shift the spectrum to the interval $[\mu, \infty)$ on the positive real axis.

78.4 An inverse source problem for Poisson's equation

We now pass to an inverse problem for Poisson's equation (78.19) assuming Ω, Γ and κ to be known: Given $u(x) = \hat{U}(x)$ for $x \in \Gamma$, find $f(x)$ for $x \in \Omega$.

We approach this problem directly in discrete form as the following least squares problem: Find $F \in V_h$ which minimizes

$$J(F) = \|U - \hat{U}\|_\Gamma^2 + \mu\|F\|_\Omega^2 \tag{78.12}$$

over V_h, where $U \in V_h$ satisifies

$$(\nabla U, \nabla v)_\Omega + \int_\Gamma \kappa U v \, ds = (F, v) \quad \text{for all } v \in V_h, \tag{78.13}$$

and V_h is the space of continuous piecewise linear functions on a given triangulation of Ω of mesh size $h(x)$. As above $\mu \geq 0$ acts as a *regularization parameter* which helps to cope with the ill-posed nature of the problem. Further, $\|\cdot\|_\Omega$ and $(\cdot,\cdot)_\Omega$ denote the $L_2(\Omega)$ norm and scalar product, and similarly, $\|\cdot\|_\Gamma$ and $(\cdot,\cdot)_\Gamma$ denote the $L_2(\Gamma)$ norm and scalar product.

We reformulate (78.12) by introducing the solution operator $B_h : V_h \to W_h$ defined by $B_h F = U_F$ on Γ, where $U_F \in V_h$ solves (78.13), and W_h is the restriction of the space V_h to the boundary Γ, that is a set of piecewise linear functions on Γ. By definition, $U_F \in V_h$ satsifies:

$$(\nabla U_F, \nabla v) + \int_\Gamma \kappa U_F \, v \, ds = (F, v) \quad \text{for all } v \in V_h. \tag{78.14}$$

We can now formulate the minimization problem (78.12) as follows: Find $F \in V_h$ which minimizes

$$J(F) = \|B_h F - \hat{U}\|_\Gamma^2 + \mu\|F\|_\Omega^2, \tag{78.15}$$

over V_h. This is a quadratic minimization problem with unique solution $F \in V_h$ characterized by a least squares equation of the form

$$(B_h F, B_h G)_\Gamma + (\mu F, G)_\Omega = (\hat{U}, B_h G)_\Omega \quad \text{for all } G \in V_h, \tag{78.16}$$

which expresses that $\frac{d}{d\epsilon} J(F + \epsilon G) = 0$ for $\epsilon = 0$ for all $G \in V_h$.

We can express (78.24) as

$$(B_h^\top B_h F, G)_\Gamma + (\mu F, G)_\Omega = (B_h^\top \hat{U}, G)_\Omega \quad \text{for all } G \in V_h,$$

that is

$$(B_h^\top B_h + \mu I) F = B_h^\top \hat{U}, \tag{78.17}$$

where $B_h^\top : W_h \to V_h$ is the transpose of B_h defined as follows: Given $w \in W_h$, we let $B_h^\top w \in V_h$ satisfy

$$(\nabla v, \nabla B_h^\top w)_\Omega + \int_\Gamma \kappa v B_h^\top w \, ds = (v, w)_\Gamma \quad \text{for all } v \in V_h. \tag{78.18}$$

In other words, $B_h^\top w$ is an approximation of the solution z of the Poisson-problem:

$$\begin{cases} -\Delta z = 0 & \text{in } \Omega, \\ \partial_n z + \kappa z = w & \text{on } \Gamma. \end{cases} \tag{78.19}$$

Choosing $v = B_h G$ in (78.18), we get using also (**??**) with $v = B_h^\top w$

$$(B_h G, w)_\Gamma = (\nabla B_h G, \nabla B_h^\top w)_\Omega + \int_\Gamma \kappa B_h G B_h^\top w \, ds = (G, B_h^\top w)_\Omega$$

and thus as expected from a transpose

$$(B_h G, w)_\Gamma = (G, B_h^\top w)_\Omega;$$

Moving B_h from G onto w brings in the transpose $B_h^\top$.

Solving (78.24) gives an approximation $F(x)$ of the function $f(x)$ we are looking for. We may solve (78.24) by direct matrix inversion if the number of nodes is small, and by some iterative method such as the gradient or the conjugate gradient method for larger problems.

The gradient method takes the form:

$$F^{n+1} = F^n - \alpha(B_h^\top B_h + \mu I) F^n - B_h^\top \hat{U}) = F^n - \alpha(B_h^\top (B_h F^n - \hat{U}) + \mu F^n).$$

In each step we have to compute $B_h F$ and $B_h^\top (B_h F - \hat{U})$ corresponding to solving two Poisson problems.

Example 78.3. Kenneth! In our first application realized using Matlab we interprete (**??**) as a matrix equation explicitely formed by computing the inverses of the stiffness matrices for the problem (78.14) and the adjoint (78.18), and we then solve this matrix equation to get the nodal values of $F(x)$. May handle a couple of hundreds of nodes this way. Results: dependence on κ and μ.

EXAMPLE 78.4. Kenneth! Size of κ and μ. Gradient and conj gardient methods.

78.5 An inverse boundary value problem for Laplace's equation

Let Ω be a domain in $\mathbb{R}^2$ with boundary Γ composed of two parts Γ_1 and Γ_2. For a given function f defined on Γ_2, let u_f be the solution to the boundary value problem

$$\begin{cases} -\Delta u_f &= 0 \quad \text{in } \Omega, \\ u_f &= 0 \quad \text{on } \Gamma_1, \quad u_f = f \text{on } \Gamma_2, \end{cases} \tag{78.20}$$

and define $Bf = \frac{\partial u_f}{\partial n}$ on Γ_1, where n is the unit outward normal to Γ_1. We may think of u_f as a stationary temperature defined in Ω satisfying given boundary conditions on Γ_1 $(= 0)$ and on Γ_2 $(= f)$, and with Bf representing the heat flux on Γ_1. Suppose now we can measure the heat flux on Γ_1 and that we want to determine the temperature f on Γ_2. We thus have a situation where we have access the boundary Γ_1 and may measure both temperature $(= 0)$ and heat flux, say $\bar{q}$, and we want to determine the temperature f on the inaccessible boundary part Γ_2. This problem arises in EKG with u being a potential and Γ_1 representing the surface of the chest and Γ_2 that of the heart, and the inverse problem being to reconstruct the potential on the heart from measurements on the chest.

We formulate the reconstruction problem as a least squares optimization problem of the form: Find f on Γ_2 which minimizes

$$J(f) = \|Bf - \bar{q}\|_{\Gamma_1}^2 + \mu\|f\|_{\Gamma_2}^2,$$

where we use the notation of the previous section, and $\mu > 0$. The optimality equation as usual takes the form

$$(B^\top B + \mu I) = B^\top \bar{q},$$

where $B^\top g = \frac{\partial u^g}{\partial n}$ on Γ_2 and u^g solves the problem

$$\begin{cases} -\Delta u^g &= 0 \quad \text{in } \Omega, \\ u^g &= g \quad \text{on } \Gamma_1, \quad u^g = 0\text{on } \Gamma_2. \end{cases} \tag{78.21}$$

This is because by integrations by parts

$$(g, Bf)_{\Gamma_1} = \int_\Omega \nabla u^g \nabla u_f \, dx = (B^\top g, f)_{\Gamma_2}$$

EXAMPLE 78.5. Kenneth!!

78.6 The Backward heat equation

Another basic inverse problem is the The *Backward heat equation*: Given the temperature at final time $t = T$, find the temperature at initial time $t = 0$.

We consider this problem in the following setting: let $f(x)$ be an initial temperature and let $u(x,t)$ be the corresponding solution of the heat equation:

$$\begin{cases} \dot{u} - \Delta u = 0 & \text{in } \Omega \times (0,T], \\ \partial_n u + \kappa u = 0 & \text{on } \Gamma \times (0,T], \\ u(x,0) = f(x) & \text{for } x \in \Omega, \end{cases}$$

where the domain $\Omega \in \mathbb{R}^d$ and the coefficient $\kappa \geq 0$ are given. We consider the following inverse problem: Given the final temperature $u(x,T)$, find the initial temperature $u(x,0) = f(x)$. This corresponds to solving the heat equation "backwards".

We consider the following discrete analog of (78.6) with discretization in space: Let V_h be the usual space of continuous piecewise linear functions on a triangualtion aof Ω with mesh size $h(x)$, and let $F \in V_h$ and let $U(t) \in V_h$ be the solution of the discrete heat equation

$$\begin{cases} (\dot{U}, v)_\Omega + (\nabla U(t), \nabla v)_\Omega = 0 & \text{for } t \in (0,T],\, v \in V_h, \\ (U(0), v) = (F, v) & \text{for } v \in V_h. \end{cases} \tag{78.22}$$

The discrete inverse problem, corresponding to a discrete "backwards" heat equation, reads: Given the final temperature $U(T) = \hat{U} \in V_h$, find the initial temperature $U(0) = F \in V_h$.

To formulate this problem as a regularized least squares problem, we introduce the solution operator $B_h : V_h \to V_h$ defined as follows: $B_h F = U(T) \in V_h$, where U solves with $U(0) = F \in V_h$. The operator B_h thus takes an initial temperature F to a corresponding final temperature $B_h F$. The regularized least squares problem is now the same as that above, that is, we seek $F \in V_h$ which minimizes

$$J(F) = \|B_h F - \hat{U}\|_\Omega^2 + \mu \|F\|_\Omega^2, \tag{78.23}$$

over V_h. The unique solution $F \in V_h$ to this quadratic minimization problem is characterized by a least squares equation of the form

$$(B_h F, B_h G)_\Omega + (\mu F, G)_\Omega = (\hat{U}, B_h G)_\Omega \quad \text{for all } G \in V_h, \tag{78.24}$$

which we can express (78.24) as

$$(B_h^\top B_h F, G)_\Omega + (\mu F, G)_\Omega = (B_h^\top \hat{U}, G)_\Omega \quad \text{for all } G \in V_h,$$

that is

$$(B_h^\top B_h + \mu I) F = B_h^\top \hat{U}, \tag{78.25}$$

where $B_h^\top : V_h \to V_h$ is defined as follows: $B_h^\top G = Z(0) \in V_h$, where $Z(t) \in V_h$ solves the discrete heat equation

$$\begin{cases} -(v, \dot{Z})_\Omega + (\nabla v, \nabla Z)_\Omega = 0 & \text{for } t \in (o, T],\ v \in V_h, \\ (v, Z(T))_\Omega = (G, v)_\Omega & \text{for } v \in V_h. \end{cases} \tag{78.26}$$

Note computing $B_h^\top$ corresponds to solving a heat equation: note the minus sign in the term $-(v, \dot{Z})$ and that we solve starting with $t = T$ and ending with $t = 0$. Changing variables introducing a new time variable $s = T - t$ brings this problem into the form of the usual heat equation. Note that is a discrete analog of the problem:

$$\begin{cases} -\dot{z} - \Delta z = 0 & \text{in } \Omega \times (0, T], \\ \partial_n z + \kappa z = 0 & \text{on } \Gamma \times [0, T), \\ z(x, T) = g(x) & \text{for } x \in \Omega, \end{cases}$$

with $G \in V_h$ and approximation of $g(x)$ and $Z(t) \in V_h$ and approximation of $z(\cot, t)$ for $t \in [0, T]$.

To solve the least squares equation by the gradient or conjugate gradient method, we have to compute $B_h F^n$ and $B_h^\top G$ for given vectors F^n and G in V_h, by using some time stepping method such as the dG(0) or cG(1) method.

Example 78.6. Kenneth. Role of α

78.7 Further applications

Wealth of inverse problems:.....

79
Differential Equations Tool Bag

It seems to me that there are at least four different viewpoints– or extremes of viewpoint– that one may reasonably hold:

1. All thinking is computation; in particular, feelings of conscious awareness are evoked merely by the carrying out of appropriate computations. (Hard AI)
2. Awareness is a feature of the brain's physiological action; and whereas any physical action can be simulated computationally, computational simulation cannot by itself evoke awareness. (Soft AI)
3. Appropriate physical action of the brain evokes awareness, but this physical action cannot even be properly simulated computationally. (Pensrose's view)
4. Awareness cannot be explained by physical, computational, or any other scientific terms.

(R. Penrose in *Shadows of the Mind*)

79.1 Introduction

We here collect basic facts about solving differential equations analytically and numerically.

79.2 The equation $u'(x) = \lambda(x)u(x)$

The solution to the scalar intial value problem

$$u'(x) = \lambda(x)u(x) \quad \text{for } x > a, \quad u(a) = u_a,$$

where $\lambda(x)$ is a given function of x, and u_a a given initial value, is

$$u(x) = \exp(\Lambda(x))u_a = e^{\Lambda(x)}u_a,$$

where $\Lambda(x)$ is a primitive function of $\lambda(x)$ such that $\Lambda(a) = 0$. In particular, if λ is a constant, then $u(x) = \exp(\lambda x)u_a$.

79.3 The equation $u'(x) = \lambda(x)u(x) + f(x)$

The solution the scalar intial value problem

$$u'(x) = \lambda(x)u(x) + f(x) \quad \text{for } x > a, \quad u(a) = u_a,$$

where $\lambda(x)$ and $f(x)$ are given functions of x, and u_a a given initial value, can be expressed using Duhamel's principle in the form

$$u(x) = e^{\Lambda(x)}u_a + e^{\Lambda(x)} \int_a^x e^{-\Lambda(y)} f(y)\, dy.$$

where $\Lambda(x)$ is a primitive function of $\lambda(x)$ such that $\Lambda(a) = 0$.

79.4 The differential equation $\sum_{k=0}^n a_k D^k u(x) = 0$

A solution to the constant coefficient differential equation

$$p(D)u(x) = \sum_{k=0}^{n} a_k D^k u(x) = 0, \quad \text{for } x \in I,$$

where I is an interval of real numbers, has the form

$$u(x) = \alpha_1 \exp(\lambda_1) + ... + \alpha_n \exp(\lambda_n),$$

where the α_i are arbitrary constants and the λ_i are the roots of the polynomial equation $p(\lambda) = 0$ with $p(\lambda) = \sum_{k=0}^n a_k \lambda^k$, assuming there are n distinct roots. If $p(\lambda) = 0$ has a multiple root λ_i of multiplicity r, then the solution is the sum of terms of the form form $q(x)\exp(\lambda_i x)$, where $q(x)$ is a polynomial of degree at most $r-1$. For example, if $p(D) = (D-1)^2$, then a solution of $p(D)u = 0$ has the form $u(x) = (a_0 + a_1 x)\exp(x)$.

79.5 The damped linear oscillator

A solution $u(t)$ to

$$\ddot{u} + \mu\dot{u} + ku = 0, \text{for } t > 0,$$

where μ and k are constants, has the form

$$u(t) = ae^{-\frac{1}{2}(\mu+\sqrt{\mu^2-4k})t} + be^{-\frac{1}{2}(\mu-\sqrt{\mu^2-4k})t},$$

if $\mu^2 - 4k > 0$, and

$$u(t) = ae^{-\frac{1}{2}\mu t}\cos(\frac{t}{2}\sqrt{4k-\mu^2}) + be^{-\frac{1}{2}\mu t}\sin(\frac{t}{2}\sqrt{4k-\mu^2}),$$

if $\mu^2 - 4k < 0$, and

$$u(t) = (a + bt)e^{-\frac{1}{2}\mu t},$$

if $\mu^2 - 4k = 0$, where a and b are arbitrary constants.

79.6 The matrix exponential

The solution to the initial value problem linear system

$$u'(x) = Au(x) \quad \text{for } 0 < x \leq T,\ u(0) = u_0,$$

where A is a *constant* $d \times d$ matrix, $u_0 \in \mathbb{R}^d$, $T > 0$, is given by

$$u(x) = \exp(xA)u_0 = e^{xA}u_0.$$

If A is diagonalizable so that $A = SDS^{-1}$, where S is nonsingular and D is diagonal with diagonal elements d_i (the eigenvalues of A), then

$$\exp(xA) = S\exp(xD)S^{-1}.$$

where $\exp(xD)$ be the diagonal matrix with diagonal elements equal to $\exp(xd_i)$.

The solution to the initial value problem

$$u'(x) = Au(x) + f(x) \quad \text{for } 0 < x \leq 1,\ u(0) = u_0,$$

where $f(x)$ is a given function, is given by Duhamel's principle:

$$u(x) = \exp(xA)u_0 + \int_0^x \exp((x-y)A)f(y)\,dy.$$

79.7 Fundamental solutions of the Laplacian

The function $\Phi(x) = \frac{1}{4\pi}\frac{1}{\|x\|}$ for $x \in \mathbb{R}^3$ satsifies the differential equation $-\Delta\Phi = \delta_0$ in $\mathbb{R}^3$, where δ_0 represents a point mass at the origin. The function $\Phi(x) = \frac{1}{2\pi}\log(\frac{1}{\|x\|})$ for $x \in \mathbb{R}^2$ satsifies the differential equation $-\Delta\Phi = \delta_0$ in $\mathbb{R}^2$, where δ_0 represents a point mass at the origin.

79.8 The wave equation in 1d

The general solution to the one-dimensional wave equation

$$\ddot{u} - u'' = 0 \quad \text{for } x, t \in \mathbb{R},$$

is given by $u(x,t) = v(x-t) + w(x+t)$ where $v, w : \mathbb{R} \to \mathbb{R}$ are arbitrary functions.

79.9 Numerical methods for initial value problems

The dG(O), the discontinuous Galerkin method with piecewise constants, for the initial value problem $\dot{u}(t) = f(u(t),t)$ for $t > 0$, $u(0) = u^0$, with $f : \mathbb{R}^{d+1} \to \mathbb{R}^d$, takes the form

$$U^n = U^{n-1} + \int_{t_{n-1}}^{t_n} f(U^n, t)\, dt, \quad n = 1, 2, ...,$$

where $U(t)$ is piecewise constant on a partition $0 = t_0 < t_1 < \cdots < t_n < t_{n+1} < \cdots$, with $U(t) = U^n$ for $t \in (t_{n-1}, t_n]$ and $U(0) = u^0$. With right-end point quadrature we obtain the implicit backward-Euler method:

$$U^n = U^{n-1} + k_n f(U^n, t_n)\, dt, \quad n = 1, 2, ...,$$

where $k_n = t_n - t_{n-1}$. The explicit forward Euler method reads:

$$U^n = U^{n-1} + k_n f(U^{n-1}, t_{n-1})\, dt, \quad n = 1, 2, ...,$$

The cG(1), the continuous Galerkin method with continuous piecewise linear functions, takes the form

$$U(t_n) = U(t_{n-1}) + \int_{t_{n-1}}^{t_n} f(U(t), t)\, dt, \quad n = 1, 2, ...,$$

where $U(t)$ is continuous piecewise linear with nodal values $U(t_n) \in \mathbb{R}^d$ and $U(0) = u^0$.

79.10 cg(1) for convection-diffusion-reaction

The cG(1) finite element method for the scalar convection-diffusion-reaction problem

$$\begin{aligned} -\nabla\cdot(a\nabla u)+\nabla\cdot(ub)+cu &= f \quad \text{in } \Omega, \\ a\frac{\partial u}{\partial n}+\kappa u &= g \quad \text{on } \Gamma, \end{aligned}$$

with Robin boundary conditions, where f and g are given data, and $a > 0$, b, c and $\kappa \geq 0$ are given coefficients, and Ω is a given domain in $\mathbb{R}^2$ with boundary Γ, takes the form: Find $U \in V_h$ such that

$$\int_\Omega a\nabla U\cdot\nabla v\,dx+\int_\Omega \nabla\cdot(ub)v\,dx+\int_\Omega cuv\,dx+\int_\Gamma \kappa uv\,ds=\int_\Omega fv\,dx+\int_\Gamma gv\,ds,$$

where V_h is a space of continuous piecewise linear functions on a triangulation of Ω with no restriction on the nodal values on the boundary.

79.11 Svensson's Formula for Laplace's equation

$$U_{i,j}=\frac{1}{4}(U_{i-1,j}+U_{i+1,j}+U_{i,j-1}+U_{i,j+1}), \quad \text{for } i,j\in\mathbb{Z},$$

where $U_{i,j}$ approximates $u(ih,jh)$ with $h > 0$ and $u : \mathbb{R}^2 \to \mathbb{R}$ solves $\Delta u = 0$.

80
Applications Tool Bag

80.1 Introduction

In this section we collect the basic models of engineering and science expressed as differential equations. For specification of boundary and initial values we refer to the text.

80.2 Malthus' population model:

$$\dot{u} = \lambda u - \mu u,$$

where $u(t)$ is the population at time t, $\lambda \geq 0$ the birth rate and $\mu \geq 0$ the death rate.

80.3 The logistics equation

$$\dot{u} = u(1 - u)$$

80.4 Mass-spring-dashpot system

$$m\ddot{u} + \mu\dot{u} + ku = f, \quad \text{((force balance)},$$

where $u(t)$ is the displacement, m is the mass, μ the viscosity, and k the spring constant.

80.5 LCR-circuit

$$L\ddot{u} + R\dot{u} + \frac{u}{C} = f, \quad ((\text{balance of potentials}),$$

where $u(t)$ is a primitive function of the current, L is the inductance, R the resistance, C the capacitance, and f a potential.

80.6 Laplace's equation for gravitation:

$$-\Delta u = \rho,$$

where $u : \mathbb{R}^3 \to \mathbb{R}$ is a gravitational potential and $\rho(x)$ is the mass density.

80.7 The heat equation:

$$\dot{u} - \nabla \cdot q = f, \quad q = k\nabla u \quad (\text{heat balance and Fourier's law})$$

where $u(x,t)$ is a temperature, $q(x,t)$ a heat flux, $k(x,t) > 0$ a conduction coefficient and $f(x,t)$ a heat source. If $k = 1$, then we get the heat equation: $\dot{u} - \Delta u = f$.

80.8 The wave equation:

$$\ddot{u} - \Delta u = f.$$

80.9 Convection-diffusion-reaction:

$$\dot{u} + \nabla \cdot (\beta u) + \alpha u - \nabla \cdot (\epsilon \nabla u) = f.$$

where $u(x,t)$ a concentration, $\beta(x,t)$ is a convection velocity, $\alpha(x,t)$ a reaction coefficient, $\epsilon(x,t)$ a diffusion coefficient, and $f(x,t)$ a production rate.

.

80.10 Maxwell's equations:

$$\begin{cases} \dfrac{\partial B}{\partial t} + \nabla \times E = 0, & \text{(Faraday's law)} \\ -\dfrac{\partial D}{\partial t} + \nabla \times H = J, & \text{(Ampère's law)} \\ \nabla \cdot B = 0, \quad \nabla \cdot D = \rho, & \text{(Gauss' and Coulomb's laws} \\ B = \mu H, \quad D = \epsilon E, \quad J = \sigma E, & \text{(constitutive laws and Ohm's law)} \end{cases}$$

where E is the electric field, H is the magnetic field , D is the electric displacement, B is the em magnetic flux , J is the electric current, ρ is the charge, μ is the magnetic permeability , ϵ is the dielectric constant, and σ is the electric conductivity.

80.11 The incompressible Navier-Stokes equations:

$$\frac{\partial u}{\partial t} + (u \cdot \nabla)u + \nabla p - \nu \Delta u = f, \quad \nabla \cdot u = 0,$$

where $u(x,t)$ is the fluid velocity, $p(x,t)$ the pressure, $f(x,t)$ a given force and $\nu > 0$ a constant viscosity.

81
Analytic functions

> A mathematician of the first rank, Laplace quickly revealed himself as only a mediocre administrator, from his first work we saw we had been deceived. Laplace saw no question from its true point of view, he sought subtleties everywhere, had only doubtful ideas, and finally carried the spirit of the infinitely small into administration. (Napoleon)
>
> We arrive at truth, not by reason only, but also by the heart. (Pascal)

81.1 Introduction

In this chapter we give a short account of *analytic functions*, that is, differentiable functions $f : \mathbb{C} \to \mathbb{C}$. We use heavily the material developed above on Calculus in $\mathbb{R}^d$, $d = 1, 2$, including, the definition of derivative of a function $f : \mathbb{R}^d \to \mathbb{R}^d$, and Green's formulas in $\mathbb{R}^2$.

81.2 The definition of an analytic function

We recall that we can write each complex number $z \in \mathbb{C}$ in the form $z = x + iy$, with $x, y \in \mathbb{R}$ and i the imaginary unit, and we can identify $\mathbb{C}$ with $\mathbb{R}^2$ by identifying $x + iy \in \mathbb{C}$ with $(x, y) \in \mathbb{R}^2$. In particular, $i = (0, 1)$, and $|z| = (x^2 + y^2)^{1/2}$.

Let $f : \Omega \to \mathbb{C}$ be a complex-valued function of a complex variable $z = x + iy \in \Omega$, where $x, y \in \mathbb{R}$ and Ω is an open domain of the complex plane. Decomposing into real and imaginary parts, we can write

$$f(z) = f(x+iy) = u(x,y) + iv(x,y),$$

where $u : \mathbb{R}^2 \to \mathbb{R}$ and $v : \mathbb{R}^2 \to \mathbb{R}$ are the real and imaginary parts of $f(z)$, that is, $u(x,y) = \text{Re } f(z)$ and $v(x,y) = \text{Im } f(z)$, where we thus view $u(x,y)$ and $v(x,y)$ as functions of $(x,y) \in \mathbb{R}^2$ with values in $\mathbb{R}$.

We say that $f : \Omega \to \mathbb{C}$ is *differentiable* at $z_0 \in \Omega$ with derivative $f'(z_0) \in \mathbb{C}$, if for z close to z_0, we have

$$|f(z) - f(z_0) - f'(z_0)(z - z_0)| \leq K_f(z_0)|z - z_0|^2, \qquad (81.1)$$

where $K_f(z_0)$ is a non-negative real constant depending on f and z_0. This is a direct extension of the corresponding definition of the derivative of a function $f : \mathbb{R} \to \mathbb{R}$ to a function $f : \mathbb{C} \to \mathbb{C}$, and the usual rules for differentiation of sums, products and quotients directly extend.

We recall that differentiability of a function $f : \mathbb{R} \to \mathbb{R}$ at a point x_0 means that $f(x)$ is well approximated (up to a quadratic term) by the linear function $c_0 + c_1(x - x_0) = f(x_0) + f'(x_0)(x - x_0)$ for x close to x_0, where c_0 and $c_1 = f'(x_0)$ are real constants. Similarly, differentiability of a function $f : \mathbb{C} \to \mathbb{C}$ at a point z_0 means that $f(z)$ for z close to z_0 is well approximated by the linear function $c_0 + c_1(z - z_0) = f(z_0) + f'(z_0)(z - z_0)$, involving a translation and multiplication by a complex constant. We conclude that differentiability of a function $f : \mathbb{C} \to \mathbb{C}$ at a point z_0 means that $f(z)$ in a neighborhood of z_0 acts like a combination of a translation, rotation and change of modulus, see Fig. **??**

Fix this fig

We say that $f : \Omega \to \mathbb{C}$ is *analytic* in the open domain Ω of the complex plane if $f(z)$ is differentiable at all $z_0 \in \Omega$ with derivative $f'(z_0)$. The derivative f' of an analytic function $f : \Omega \to \mathbb{C}$ is again a function $f' : \Omega \to \mathbb{C}$. We shall shortly prove the surprising fact that if $f : \Omega \to \mathbb{C}$ is analytic, then also $f' : \Omega \to \mathbb{C}$ is analytic with derivative $f'' : \Omega \to \mathbb{C}$, which is also analytic, and so on. An analytic function $f : \Omega \to \mathbb{C}$ thus has derivatives of all orders $f^{(n)} : \Omega \to \mathbb{C}$, $n = 1, 2, ...$, which are all analytic. We recall that a differentiable function $f : \mathbb{R} \to \mathbb{R}$ need not have a differentiable derivative, and therefore does not have this very special property in general.

We can view a function $f : \mathbb{C} \to \mathbb{C}$ alternatively as a function $f : \mathbb{R}^2 \to \mathbb{R}^2$ if we identify $\mathbb{C}$ and $\mathbb{R}^2$ as indicated. The Jacobian $f'(x,y)$ of a function $f : \mathbb{R}^2 \to \mathbb{R}^2$ is a 2×2-matrix consisting of 4 real numbers, while the derivative $f'(z) \in \mathbb{C}$ of a function $f : \mathbb{C} \to \mathbb{C}$ is supposed to be a complex number being represented by 2 real numbers. We conclude that differentiablity of a complex-valued function $f : \mathbb{C} \to \mathbb{C}$, is a more stringent requirement than differentiablity of the corresponding function $f : \mathbb{R}^2 \to \mathbb{R}^2$, which only requires the partial derivatives of the real and imaginary parts $u(x,y)$ and $v(x,y)$ of $f(z)$ to exist. In fact, we shall see that the partial derivatives of the real and imaginary parts of an analytic function must be coupled in a specific way, which is expressed through the *Cauchy-Riemann equations* stated below.

81.3 The derivative as a limit of difference quotients

Note that (84.1) implies that if $z \neq z_0$, then

$$|\frac{f(z) - f(z_0)}{z - z_0} - f'(z_0)| \leq K_f(z_0)|z - z_0|,$$

which we can write in the form

$$\lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} = f'(z_0), \tag{81.2}$$

by which we mean, of course, that

$$|\frac{f(z) - f(z_0)}{z - z_0} - f'(z_0)|$$

is a small as we please if we only choose $|z - z_0|$ small enough (respecting that $z \neq z_0$). In view of (81.2) we write as usual $\frac{df}{dz} = f'$.

81.4 The function $f(z) = az + b$ is analytic in $\mathbb{C}$

We consider the function $f : \mathbb{C} \to \mathbb{C}$ given by $f(z) = az + b$, where a and b are given complex numbers. We have for all z and $z_0 \in \mathbb{C}$ that

$$f(z) - f(z_0) - a(z - z_0) = 0,$$

and thus $f(z)$ is analytic in $\mathbb{C}$ with derivative $f'(z) = a$.

81.5 The function $f(z) = z^2$ is analytic in $\mathbb{C}$

If $f(z) = z^2$, then

$$f(z) - f(z_0) - 2z_0(z - z_0) = z^2 - z_0^2 - 2z_0 z + 2z_0^2 = (z - z_0)^2,$$

and thus $f'(z_0) = 2z_0$ for $z_0 \in \mathbb{C}$.

81.6 The function $f(z) = z^n$ is analytic in $\mathbb{C}$ for $n = 1, 2, ...$

Consider the function $f : \mathbb{C} \to \mathbb{C}$, where $f(z) = z^n$ and $n = 1, 2, ...$, is a natural number, which may be viewed as an extension of the function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^n$. One can prove by a direct extension of the proof in the case $f(x) = x^n$, that

$$f'(z) = nz^{n-1}. \tag{81.3}$$

We conclude that z^n is differentiable in the whole of $\mathbb{C}$, with derivative nz^{n-1}. We just gave the proof in the case $n = 1, 2$.

81.7 Rules of differentiation

As we said, the usual rules for differentiation of sums, products and quotients valid for functions $f : \mathbb{R} \to \mathbb{R}$ extend to functions $f : \mathbb{C} \to \mathbb{C}$. In particular we have if $f(z_0) \neq 0$ and $g(z) = \frac{1}{f(z)}$, that

$$g'(z_0) = -\frac{f'(z_0)}{f^2(z_0)}. \tag{81.4}$$

Further, the composition of two analytic functions is also analytic and the Chain rule for differentiation holds: if $g(z)$ is differentiable at z_0 and $f(z)$ is differentiable at $g(z_0)$, then the composite function $h(z) = f(g(z))$ is differentiable at z_0 with derivative $h'(z_0) = f'(g(z_0))g'(z_0)$. The proof is a direct extension of the corresponding proof for real valued functions of a real variable.

81.8 The function $f(z) = z^{-n}$ is analytic for $z \neq 0$ if $n = 1, 2, ...$

Applying the rule (81.4), we see that $f(z) = z^{-n}$ with $n = 1, 2, ...$, is differentiable for $z \neq 0$ and

$$f'(z) = -nz^{-n-1}, \quad \text{for } z \neq 0. \tag{81.5}$$

We can summarize by stating that if $f(z) = z^n$ with $n = \pm 1, \pm 2, ...$, then $f'(z) = nz^{n-1}$, where we assume that $z \neq 0$ if $n < 0$.

81.9 The Cauchy-Riemann equations

We shall now derive the so-called *Cauchy-Riemann equations* connecting the partial derivatives of the real part $u(x, y)$ and the imaginary part $v(x, y)$

of a complex-valued function $f(z) = u(x,y)+iv(x,y)$ of a complex variable $z = x + iy$ at a point $z_0 = x_0 + iy_0$ with $x_0, y_0 \in \mathbb{R}$, such that (84.1) is satisfied. Writing $f'(z_0) = a + ib$, with $a, b \in \mathbb{R}$, we can express (84.1) as

$$\begin{aligned}&|u(x,y) + iv(x,y) - u(x_0,y_0) - iv(x_0,y_0) - (a+ib)(x - x_0 + i(y - y_0))| \\ &\leq K_f(z_0)|z - z_0|^2.\end{aligned}$$

Separating into real and imaginary parts, we conclude (recalling (22.7)) that

$$\begin{aligned}|u(x,y) - u(x_0,y_0) - a(x - x_0) + b(y - y_0)| \leq K_f(z_0)|z - z_0|^2,\\ |v(x,y) - v(x_0,y_0) - a(y - y_0) - b(x - x_0)| \leq K_f(z_0)|z - z_0|^2.\end{aligned} \tag{81.6}$$

Recalling the definition of the partial derivatives of $u(x,y)$ and $v(x,y)$ at (x_0,y_0) from Chapter *Vector-valued functions of several variables*, we conclude that

$$\begin{aligned}a &= \frac{\partial u}{\partial x}(x_0,y_0), \quad b = -\frac{\partial u}{\partial y}(x_0,y_0),\\ a &= \frac{\partial v}{\partial y}(x_0,y_0), \quad b = \frac{\partial v}{\partial x}(x_0,y_0),\end{aligned}$$

and we thus find that

$$\frac{\partial u}{\partial x}(x_0,y_0) = \frac{\partial v}{\partial y}(x_0,y_0), \qquad \frac{\partial u}{\partial y}(x_0,y_0) = -\frac{\partial v}{\partial x}(x_0,y_0). \tag{81.7}$$

These are the *Cauchy-Riemann equations* for $u(x,y)$ and $v(x,y)$ at the point (x_0,y_0).

We conclude that if $f(z) = u(x,y) + iv(x,y)$ is analytic in the open domain Ω of the complex plane, then the real and imaginary parts $u(x,y)$ and $v(x,y)$ satisfy the Cauchy-Riemann equations in Ω, that is,

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}, \qquad \text{in } \Omega. \tag{81.8}$$

Note that we can write the Cauchy-Riemann equations in the form $\nabla v = -\nabla \times u$, recalling that $\nabla \times u = (\frac{\partial u}{\partial y}, -\frac{\partial u}{\partial x})$.

EXAMPLE 81.1. The analytic function $f(z) = z^2 = (x+iy)^2 = x^2-y^2+2ixy$ with $u(x,y) = x^2 - y^2$ and $v(x,y) = 2xy$ satisfies $\frac{\partial u}{\partial x} = 2x = \frac{\partial v}{\partial y}$ and $\frac{\partial u}{\partial y} = -2y = -\frac{\partial v}{\partial x}$.

We have seen that the Cauchy-Riemann equations (81.8) follow from the analyticity of the complex valued function $f = u + iv$. In other words, the Cauchy-Riemann equations represents a *necessary condition* for the analyticity of $f = u + iv$. The Cauchy-Riemann equations also represent a

FIGURE 81.1. Riemann

sufficient condition: given a pair of functions $u(x,y)$ and $v(x,y)$ satisfying the Cauchy-Riemann equations, the function $f = u + iv$ is analytic. To see this, we note that if $u(x,y)$ and $v(x,y)$ are differentiable functions satisfying (81.8), then (81.6), and thus also (84.1) holds. That is, $f = u+iv$ is analytic.

EXAMPLE 81.2. We consider the functions $u(x,y) = x + 2xy$ and $v(x,y) = y-x^2+y^2$ and find that $\frac{\partial u}{\partial x} = 1+2y = \frac{\partial v}{\partial y}$ and $\frac{\partial u}{\partial y} = 2x = -\frac{\partial v}{\partial x}$, that is, u and v satisfy the Cauchy-Riemann equations and we thus conclude that the function $f(z) = u(x,y)+iv(x,y)$ must be analytic. In fact, $f(z) = u+iv = x+2xy+i(y-x^2+y^2) = x+iy-i(x+iy)^2 = z-iz^2$, and the analyticity is obvious.

We may summarize as follows: The function $f(z) = u(x,y) = iv(x,y)$ is analytic if and only if the Cauchy-Riemann equations (81.8) are satisfied.

81.10 The Cauchy-Riemann equations and the limit definition of the derivative

Using the limit definition (81.2) of the derivative $f'(z_0)$, we can write, varying first only x:

$$f'(z_0) = \lim_{x\to x_0} \frac{f(z) - f(z_0)}{x - x_0} = \frac{\partial u}{\partial x}(x_0, y_0) + i\frac{\partial v}{\partial x}(x_0, y_0), \tag{81.9}$$

and then only y:

$$f'(z_0) = \lim_{y\to y_0} \frac{f(z) - f(z_0)}{i(y - y_0)} = \frac{1}{i}\frac{\partial u}{\partial y}(x_0, y_0) + \frac{\partial v}{\partial y}(x_0, y_0),$$

from which the Cauchy-Riemann equations follow by equating the real and imaginary parts of $f'(z_0)$ using that $\frac{1}{i} = -i$.

EXAMPLE 81.3. In the last example we found that $u(x,y) = x + 2xy$ and $v(x,y) = z - x^2 + y^2$ satisfy the Cauch-Riemann equation. According to (81.9), the derivative of the corresponding analytic function $f = u + iv$ is given by $f' = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial y} = 1 + 2y + i(-2x)$ which agrees with our observation that $f(z) = z - iz^2$ with $f'(z) = 1 - 2iz = 1 - 2i(x + iy) = 1 + 2y - 2ix$.

EXAMPLE 81.4. By direct verification using the Cauchy-Riemann equations one finds that $f(z) = e^z = e^x(\cos(y) + i\sin(y))$ is analytic in $\mathbb{C}$, and $\frac{d}{dz}e^z = e^z$. It follows that also $\sin(z) = \frac{1}{2i}(e^{iz} - e^{-iz})$ and $\cos(z) = \frac{1}{2}(e^{iz} + e^{-iz})$ are analytic in $\mathbb{C}$, and $\frac{d}{dz}\sin(z) = \cos(z)$ and $\frac{d}{dz}\cos(z) = -\sin(z)$. See Problem 81.1

81.11 The Cauchy-Riemann equations in polar coordinates

The Cauchy-Riemann equations take the following form in polar coordinates $z = re^{i\theta}$:

$$\frac{\partial u}{\partial r} = \frac{1}{r}\frac{\partial v}{\partial \theta}, \quad \frac{\partial v}{\partial r} = -\frac{1}{r}\frac{\partial u}{\partial \theta}. \tag{81.10}$$

EXAMPLE 81.5. The function Log $(z) = \log(|z|) + i\,\mathrm{Arg}\,z$ is analytic in $\{z \in \mathbb{C} : z \neq 0, 0 \leq \arg \mathrm{z} < 2\pi\}$. This follows from the Cauchy-Riemann equations in polar coordinates. We recall that $\log(z) = \log(|z|) + i\arg z$ is multi-valued since $\arg z$ is multivalued. Restricting $\arg z$ to $0 \leq \arg \mathrm{z} < 2\pi$, makes $\log(z) =$ Log (z) single-valued analytic.

81.12 The real and imaginary parts of an analytic function are harmonic

We shall now prove that the Cauchy-Riemann equations (81.8) imply that both $u(x,y)$ and $v(x,y)$ are harmonic in Ω, that is,

$$\Delta u = 0 \quad \text{and} \quad \Delta v = 0 \quad \text{in } \Omega.$$

In fact, this follows directly by differentiating (81.8) with respect to x and y, if we assume that $u(x,y)$ and $v(x,y)$ are twice differentiable, since

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 v}{\partial x \partial y} = \frac{\partial^2 v}{\partial y \partial x} = -\frac{\partial^2 u}{\partial y^2} \quad \text{in } \Omega \tag{81.11}$$

and thus $\Delta u = 0$ in Ω , and similarly $\Delta v = 0$ in Ω.

Now, one can show that solutions of the Cauchy-Riemann equations indeed must be twice differentiable, and thus the real and imaginary parts of an analytic function are harmonic. We sum up in the following theorem:

Theorem 81.1 *If $f : \Omega \to \mathbb{C}$ is analytic, where Ω is an open domain of the complex plane $\mathbb{C}$, then the real part $u(x, y) = Re\ f(z)$ and the imaginary part $v(x, y) = Im\ f(z)$ are harmonic in Ω.*

81.13 The derivative of an analytic function is an analytic function

Assume that $f(z)$ is *analytic* in the open domain Ω of the complex plane. This means that the derivative $f'(z)$ exists as a complex-valued function for $z \in \Omega$, and one may ask if $f'(z)$ itself has a derivative in Ω, that is, if $f'(z)$ is analytic in Ω?. The plain answer is YES, which we prove below. Thus, if $f(z)$ is analytic in Ω, then also $f'(z)$ is analytic in Ω, and thus also the derivative of $f'(z)$, that is the second derivative $f''(z)$ is analytic, and so on. We conclude that an analytic function has derivatives of all orders. This is a remarkable property of an analytic function.

To answer the question posed, it is sufficient to notice that if $u(x, y)$ and $v(x, y)$ satisfy the Cauchy-Riemann equations, then so do all derivatives of $u(x, y)$ and $v(x, y)$, in particular $\frac{\partial u}{\partial x}$ and $\frac{\partial v}{\partial x}$, and thus $f' = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x}$ is analytic in Ω. We state this important result as a theorem:

Theorem 81.2 *If $f : \Omega \to \mathbb{C}$ is analytic, where Ω is an open domain of the complex plane $\mathbb{C}$, then all the derivatives $f^{(n)(z)}$, $n = 1, 2, ...,$ of $f(z)$ are analytic in Ω.*

We recall that if $f : \mathbb{R} \to \mathbb{R}$ is a real-valued function of a real variable, then the analogous statement may be wrong: even if $f'(x)$ exists, it is not clear that $f''(x)$ exists.

81.14 Conjugate harmonic functions

Suppose $u(x, y)$ is harmonic in a simply connected domain Ω in $\mathbb{R}^2$. We shall now prove that there exists a harmonic function $v(x, y)$, uniquely determined up to a constant, such that $f(z) = u(x, y) + iv(x, y)$ is analytic in Ω. We say that the function $v(x, y)$ is *conjugate* to $u(x, y)$. To prove this, we simply solve the Cauchy-Riemann equations $\nabla v = -\nabla \times u$ with u given using the basic result of the Chapter Potential fields, noting that $= \nabla \times (-\nabla \times u) = \Delta u = 0$, that is, $-\nabla \times u$ is irrotational, and thus is the gradient of some function v. See also Problem 81.10.

81.15 Curves in the complex plane

Let Ω be an open domain in the complex plane $\mathbb{C}$, and let $\gamma : I \to \Omega$, where $I = [a, b]$ is an interval of $\mathbb{R}$, be a Lipschitz continuous function. We say that $\Gamma =$ Range of $\gamma = \{\gamma(t) : t \in I\}$, is a *curve* in $\mathbb{C}$ parametrized by $\gamma(t)$. For example $\gamma(t) = \exp(it)$ where $0 \leq t < 2\pi$ is a parametrization of the unit circle.

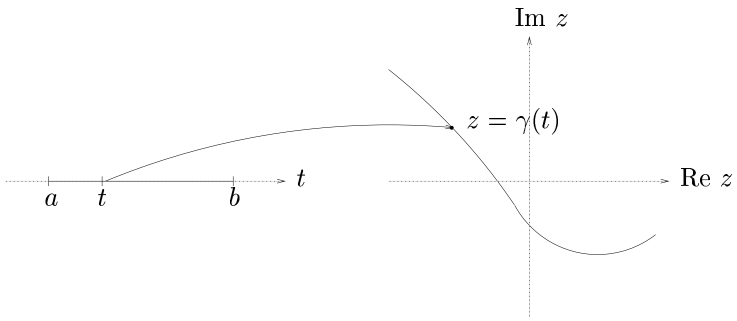


FIGURE 81.2. A curve $z = \gamma(t)$

We say that Γ is a *differentiable curve* if the corresponding parametrization $\gamma : I \to \mathbb{C}$ is differentiable on I in the sense that the related function $\gamma : I \to \mathbb{R}^2$ is differentiable. In other words, decomposing $\gamma(t) = x(t) + iy(t)$ into real and imaginary parts $x : I \to \mathbb{R}$ and $y : I \to \mathbb{R}$, we have that $\gamma(t)$ is differentiable on I if $x(t)$ and $y(t)$ are differentiable on I. We also say that $\gamma : I \to \mathbb{C}$ is Lipschitz continuous on I if the corresponding function $\gamma : I \to \mathbb{R}^2$ is Lipschitz continuous. There are thus no surprises in this context.

A curve Γ with parametrization $\gamma : [a, b] \to \mathbb{C}$ is said to *closed and simple* if $\gamma(a) = \gamma(b)$ and $\gamma(s) = \gamma(t)$ only if $a = b$.

We say that a domain Ω in $\mathbb{C}$ which is bounded by a simple closed curve, is *simply connected*. A simply connected domain does not have any "holes".

81.16 Conformal mappings

Let $f : \Omega \to \mathbb{C}$ be analytic where Ω is an open domain in $\mathbb{C}$. We shall now prove that the mapping $z \to w = f(z)$ is *conformal* in Ω in the sense that angles are preserved under the mapping $w = f(z)$. This is a direct consequence of the Chain rule and the analyticity of $f(z)$, as we now show. Let then $\gamma : I \to \mathbb{C}$, where $I = [-\delta, \delta]$ with $\delta > 0$, be a curve through $z_0 \in \Omega$ with $\gamma(0) = z_0$. Consider the curve $\kappa(t) = f(\gamma(t))$ which is the image of $\gamma(t)$ under the transformation $w = f(z)$. By the Chain rule we

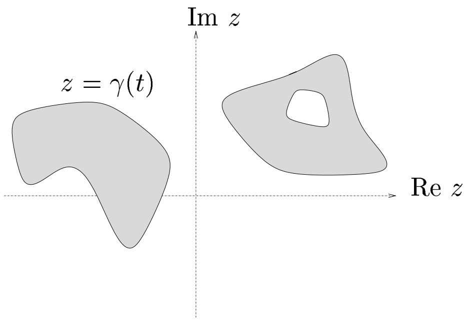


FIGURE 81.3. A simply connected domain with boundary curve $z = \gamma(t)$, to the left, and a multiply connected domain with *one* hole to the right with the boundary consisting of *two* simple closed curves

have

$$\frac{d\kappa}{dt} = \frac{df}{dz}\frac{d\gamma}{dt},$$

and we thus see, recalling that the argument of the product of two complex numbers is the sum of the arguments of the numbers, that

$$\arg\frac{d\kappa}{dt}(0) = \text{Arg } f'(z_0) + \text{Arg }\frac{d\gamma}{dt}(0),$$

where we assume that $f'(z_0) \neq 0$. Since $f(z)$ is analytic at z_0, we have that $f'(z_0)$ is independent of the curve γ, and thus the tangent direction $\frac{d\kappa}{dt}(0)$ differs from that of $\frac{d\gamma}{dt}(0)$ by the constant value Arg $f'(z_0)$, independent of γ. We conclude that the angle between two curves passing through z_0 is the same as the angle between the corresponding transformed curves passing through $f(z_0)$. This means that the mapping $w = f(z)$ is *conformal at* z_0: angles are preserved locally.

A fig here illustrating conformality with two intersecting curves γ and corresponding two intersecting curves κ

Note that since

$$\lim_{z\to z_0}\left|\frac{f(z)-f(z_0)}{z-z_0}\right| = |f'(z_0)|,$$

the mapping $w = f(z)$ changes the length scale locally by the factor $|f'(z_0)| \neq 0$. Thus although the mapping $w = f(z)$ is locally conformal, the image of a large figure may be considerably distorted because of the change of scale.

We now present some basic analytic functions $w = f(z)$ and the corresponding conformal mappings of $f : \mathbb{C} \to \mathbb{C}$.

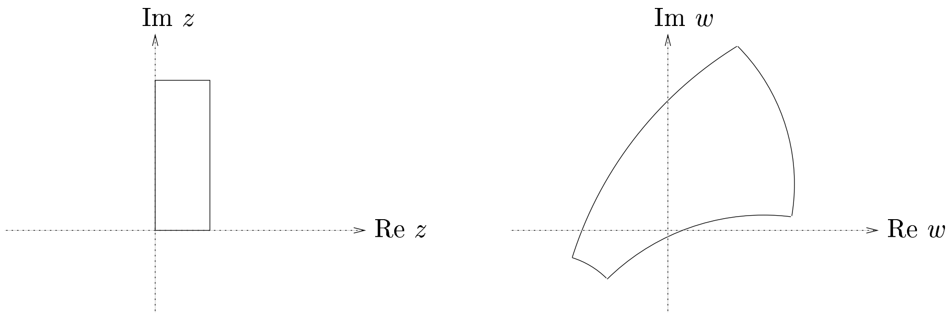


FIGURE 81.4. A conformal mapping with large deformations.

81.17 Translation-rotation-expansion/contraction

The linear transformation:

$$w = f(z) = az + b$$

where $a, b \in \mathbb{C}$, corresponds to a rotation with $\operatorname{Arg} a$ and an expansion/contraction with $|a|$, and a translation with b, see Fig. 81.5

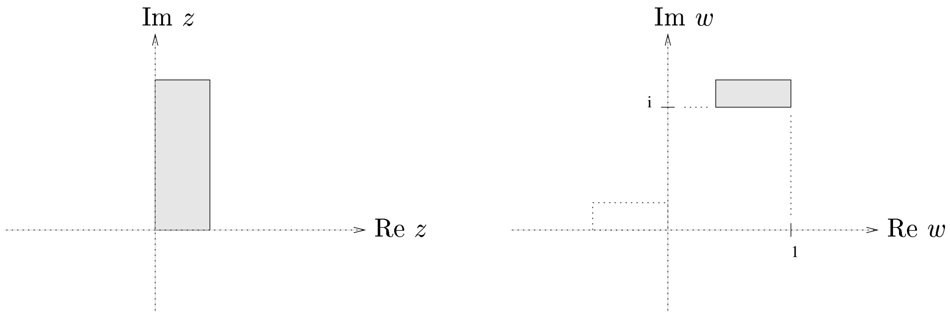


FIGURE 81.5. The mapping $w = az + b$ with $a = \frac{1}{2}i$ and $b = 1 + i$.

81.18 Inversion

The mapping

$$w = f(z) = \frac{1}{z},$$

is referred to as *inversion*. We now prove that an inversion maps every straight line or circle in the complex plane into a circle or straight line. Indeed, a circle or straight line in $\mathbb{R}^2$ can be written

$$A(x^2 + y^2) + Bx + Cy + D = 0$$

with A, B, C, D real, and $A = 0$ corresponding to a straight line. In terms of $z = x + iy$ and $\bar{z} = x - iy$, the equation takes the form

$$Az\bar{z} + B\frac{z + \bar{z}}{2} + C\frac{z - \bar{z}}{2i} + D = 0,$$

and substitution of $z = \frac{1}{w}$ gives

$$A + B\frac{\bar{w} + w}{2} + C\frac{\bar{w} - \bar{w}}{2i} + Dw\bar{w} = 0$$

which represents a circle or straight line.

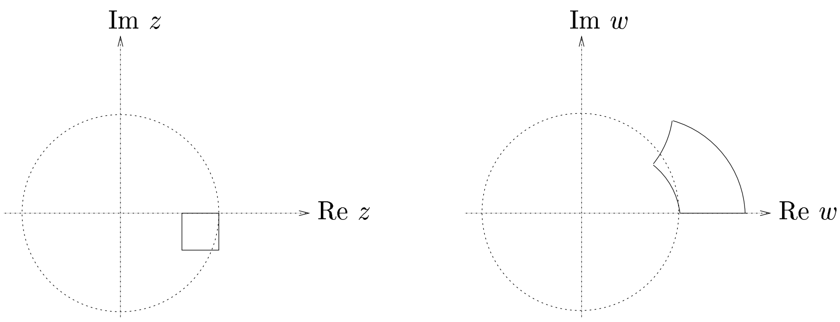


FIGURE 81.6. The mapping $w = 1/z$

81.19 Möbius transformations

A mapping of the form

$$w = f(z) = \frac{az + b}{cz + d},$$

where $a, b, c, d \in \mathbb{C}$, is said to be a *Möbius transformation*. We have

$$f'(z) = \frac{ad - bc}{(cz + d)^2},$$

and we are thus led to assume that $ad - bc \neq 0$ to guarantee conformity. Evidently, the inversion $w = \frac{1}{z}$ is a special case of a Möbius transformation. One can prove that a Möbius transformation maps every straight line or circle in the complex plane into a circle or straight line, see Problem 81.6.

EXAMPLE 81.6. (**Disc onto disc**) The function

$$w = f(z) = e^{i\alpha}\frac{z - z_0}{1 - \bar{z}_0 z}$$

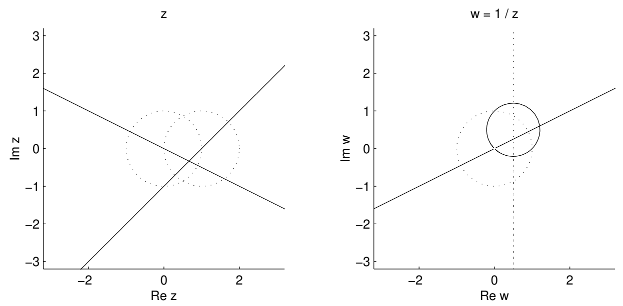


FIGURE 81.7. Further illustration of the map $f(z) = 1/z$. Note that the unit circle is mapped onto itself, while a circle through the origin is mapped onto a straight line. Note also that the straight line $y = ax$ with $a \neq 0$ passing through the origin is mapped onto its conjugate line $y = -ax$, while other lines are mapped onto circles.

where $\alpha \in \mathbb{R}$ and $z_0 \in \mathbb{C}$ with $|z_0| < 1$, maps the closed unit disc $\{|z| \leq 1\}$ onto the closed unit disc $\{|w| \leq 1\}$ in a one-to-one fashion with $f(z_0) = 0$. For the verification it suffices to verify that the circle $\{|z| = 1\}$ is mapped onto the circle $\{|w| = 1\}$.

EXAMPLE 81.7. (**Half-plane onto unit disc**) The function

$$w = f(z) = e^{i\alpha}\frac{z - z_0}{z - \bar{z}_0}$$

where $\alpha \in \mathbb{R}$ and $\text{Im } z_0 > 0$, maps the upper half-plane $\{\text{Im } z > 0\}$ onto the open unit disc $\{|w| < 1\}$ with $f(z_0) = 0$.

81.20 The functions $w = z^{1/2}$, $w = e^z$, $w = \log(z)$ and $w = \sin(z)$

We describe in a couple of examples basic aspects of the mapping properties of some elementary functions.

EXAMPLE 81.8. The function

$$w = f(z) = z^{1/2} = \sqrt{|z|}e^{\frac{i}{2}\text{Arg z}},$$

maps the wedge $\{0 \leq \arg z < \theta\}$ where $0 \leq \theta < 2\pi$ onto the wedge $\{0 \leq \arg w < \frac{\theta}{2}\}$.

EXAMPLE 81.9. The function $w = e^z$ maps the strip $\{z = x + iy : x \in \mathbb{R}, 0 \leq y < 2\pi\}$ onto the complex plane C minus the origin. The line $\{x + iy : x \in \mathbb{R}\}$ with y fixed is mapped onto the halfline $\{(r, \theta) : r > 0\}$ with $\theta = y$ using polar coordinates.

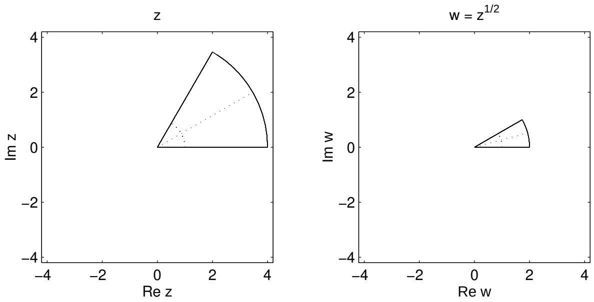


FIGURE 81.8. Illustration of the map $f(z) = z^{1/2}$.

EXAMPLE 81.10. The function $w = \mathrm{Log}(z)$ maps $\mathbb{C}$ minus the origin onto the strip $\{w \in \mathbb{C} : 0 \leq \mathrm{Im}(w) < 2\pi\}$.

EXAMPLE 81.11. The function

$$\begin{aligned} w = f(z) = \sin(z) &= \frac{1}{2i}(e^{i(x+iy)} - e^{-i(x+iy)}) \\ &= \sin(x)\cosh(y) + i\,\cos(x)\sinh(y) = u(x,y) + iv(x,y), \end{aligned}$$

maps the strip $\{z = x + iy : \frac{-\pi}{2} < x < \frac{\pi}{2},\, y \in \mathbb{R}\}$ onto $\{w = u + iv : v \neq 0 \text{ if } |u| > 1\}$, which is the whole plane minus the two half-lines $\{u + iv : |u| > 1, v = 0\}$. The level curves are hyperbolas and ellipses.

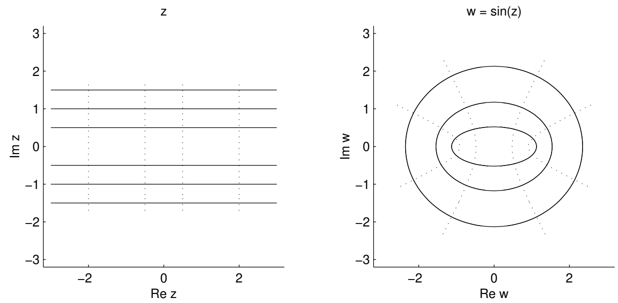


FIGURE 81.9. Illustration of the map $f(z) = sin(z)$.

81.21 Complex integrals: first shot

We make a direct extension of the integral of a differentiable function $F : \mathbb{R} \to \mathbb{R}$ to the integral of an analytic function $F : \mathbb{C} \to \mathbb{C}$, paralleling closely the presentation in Chapter *The Integral*.

Let $F(z)$ be analytic in the domain Ω of the complex plane, with Lipschitz continuous derivative $f(z) = F'(z)$. Let Γ be a differentiable curve in Ω parametrized by $\gamma : [a, b] \to \mathbb{C}$, connecting the point $z_a = \gamma(a)$ with the point $z_b = \gamma(b)$, and let $z_a = z_0, z_1, ..., z_n = z_b$ be a sequence of points on Γ connecting z_a and z_b , see Fig. 81.10. We assume that $z_k \neq z_{k-1}$ for $k = 1, ..., n$.

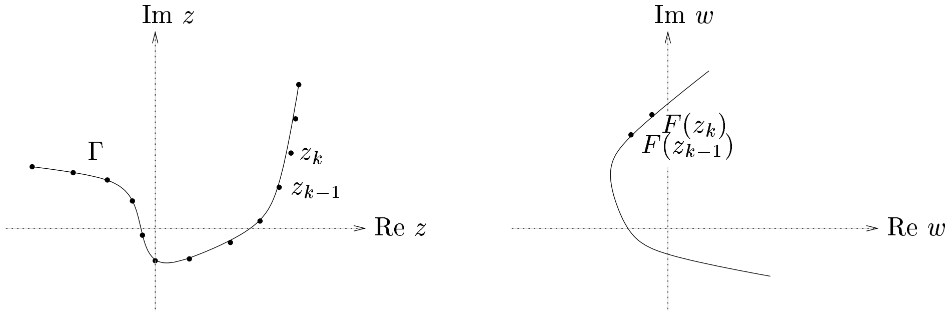


FIGURE 81.10. A curve Γ with sample points z_k and corresponding function values $F(z_k)$.

We can write

$$F(z_b) - F(z_a) = \sum_{k=1}^{n}(F(z_k) - F(z_{k-1})) = \sum_{k=1}^{n} \frac{F(z_k) - F(z_{k-1})}{z_k - z_{k-1}}(z_k - z_{k-1}). \tag{81.12}$$

Letting $\max_{k=1,\dots,n} |z_k - z_{k-1}|$ tend to zero, we are led to write

$$F(z_b) - F(z_a) = \int_\Gamma f(z)\, dz, \tag{81.13}$$

where we replace $\frac{F(z_k)-F(z_{k-1})}{z_k-z_{k-1}}$ by the derivative $F'(z_{k-1}) = f(z_{k-1})$ and $z_k - z_{k-1}$ by dz.

We note that the integral $\int_\Gamma f(z)\, dz$, being equal to $F(z_b) - F(z_a)$, is independent of the choice of the curve Γ connecting z_a and z_b. As a special case we note that if Γ is closed, corresponding to choosing $z_b = z_a$, then

$$\int_\Gamma f(z)\, dz = 0. \tag{81.14}$$

Recalling that $f(z)$ is analytic if $F(z)$ is analytic, we have found a reason to believe in *Cauchy's theorem* stating that the integral of an analytic function $f : \Omega \to \mathbb{C}$ around a simple closed curve in Ω enclosing a region contained in Ω, is zero. This is a corner-stone of the theory of analytic functions. Below we give a proof of Cauchy's theorem using a Green's formula.

81.22 Complex integrals: general case

Let Ω be an open domain in the complex plane and let Γ be a differentiable curve in $\mathbb{C}$ parametrized by $\gamma = (x, y) : [a, b] \to \mathbb{C}$. Let $f = u + iv : \Gamma \to \mathbb{C}$ be Lipschitz continuous and define

$$\int_\Gamma f(z)\,dz = \int_a^b \big(u(x(t), y(t)) + iv(x(t), y(t))\big)(\dot{x}(t) + i\dot{y}(t))\,dt, \quad (81.15)$$

where thus formally $dz = dx + idy = \dot{x}dt + i\dot{y}dt = (\dot{x} + i\dot{y})\,dt$. The integral is defined if $u(x, y)$ and $v(x, y)$ are Lipschitz continuous in (x, y) and $\dot{x}(t)$ and $\dot{y}(t)$ are Lipschitz continuous in t. As in the Chapter Curve Integrals, we see that the integral is independent of the parametrization.

We can express the integral as a limit of Riemann sums in the usual way:

$$\int_\Gamma f(z)\,dz = \lim_{n\to\infty} \sum_{k=1}^n f(z_{k-1})(z_k - z_{k-1}), \quad (81.16)$$

where $z_a = z_0, z_1, ..., z_n = z_b$ is a sequence of points along Γ with $\max_{k=1,\dots,n} |z_k - z_{k-1}|$ tending to zero as n tends to infinity.

Below we use also ζ as a complex variable and thus write in particular

$$\int_\Gamma f(z)\,dz = \int_\Gamma f(\zeta)\,d\zeta.$$

EXAMPLE 81.12. Let Γ be a circle around the origin of radius one oriented counter-clockwise and parametrized by $\gamma(t) = e^{it} = \cos(t) + i\sin(t) = (\cos(t), \sin(t))$ with $0 \le t < 2\pi$. Let $f(z) = z^n$ with n an integer. We have for $n \neq -1$, since $dz = (-\sin(t) + i\cos(t))\,dt = ie^{it}\,dt$,

$$\begin{aligned}\int_\Gamma f(z)\,dz &= \int_\Gamma z^n\,dz = \int_0^{2\pi} e^{int}(-\sin(t) + i\cos(t))\,dt = i\int_0^{2\pi} e^{int}e^{it}\,dt \\ &= i\int_0^{2\pi} e^{i(n+1)t}\,dt = \frac{i}{n+1}[\sin((n+1)t) - i\cos((n+1)t)]_0^{2\pi} = 0.\end{aligned}$$

This conforms with Cauchy's theorem for $n = 0, 1, 2, ...$, since then $f(z)$ is analytic in $\mathbb{C}$. For $n = -1$ with $f(z) = \frac{1}{z}$, we get

$$\int_\Gamma \frac{1}{z}\,dz = \int_\Gamma \frac{dz}{z} = \int_0^{2\pi} \frac{ie^{it}}{e^{it}}\,dt = 2\pi i. \quad (81.17)$$

Note the counter-clockwise orientation of Γ. The function $f(z) = \frac{1}{z}$ is not analytic in the domain enclosed by Γ, since $\frac{1}{z}$ is not differentiable for $z = 0$, and thus the integral $\int_\Gamma \frac{dz}{z}$ may be non-zero. We shall see below that the derivative of $\log(z)$ is equal to $\frac{1}{z}$, but $\log(z)$ is not uniquely defined for $z \neq 0$, and thus $\int_\Gamma \frac{dz}{z}$ may be non-zero.

The functions $f(z) = z^n$ with $n = -2, -3, ..$ are all derivatives of analytic functions and thus $\int_\Gamma f(z)\,dz = 0$ if Γ is a closed curve which does not pass through 0.

81.23 Basic properties of the complex integral

The complex integral has properties analogous to those of the usual real integral such as linearity, additivity over subintervals and integration by parts. For example, we have if $|f(z)| \leq M$ for $z \in \Gamma$:

$$|\int_\Gamma f(z)\,dz| \leq M \int_\Gamma ds = ML(\Gamma), \tag{81.18}$$

where $L(\Gamma)$ is the length of Γ:

$$L(\Gamma) = \int_a^b (\dot{x}^2(t) + \dot{y}^2(t))^{1/2}\,dt.$$

This follows by taking absolute values in (81.16) and then passing to the limit:

$$|\int_\Gamma f(z)\,dz| \leq \int_\Gamma |f(z)|\,|dz| = \int_\Gamma |f(z)|\,ds \leq ML(\Gamma),$$

where formally $|dz| = ds$, and thus the estimate may be viewed as a generalized triangle inequality.

81.24 Taylor's formula: first shot

If $f : \Omega \to \mathbb{C}$ is analytic and Γ is a straight line in Ω connecting z_a and z_b, then we can write

$$f(z) = f(z_0) + \int_\Gamma f'(\zeta)\,d\zeta = f(z_0) + \int_\Gamma f'(\zeta)\frac{d}{d\zeta}(\zeta - z_0)\,d\zeta,$$

and thus by partial integration (the usual rules hold)

$$f(z) = f(z_0) + \int_\Gamma f'(\zeta)\,d\zeta = f(z_0) + f'(z_0)(z - z_0) - \int_\Gamma f''(\zeta)(\zeta - z_0)\,d\zeta.$$

Continuing, writing $(\zeta - z_0) = \frac{1}{2}\frac{d}{d\zeta}(\zeta - z_0)^2$, we get

$$f(z) = f(z_0) + f'(z_0)(z - z_0) + \frac{f''}{2}(z_0)(z - z_0)^2 + \int_\Gamma f^{(3)}(\zeta)\frac{(\zeta - z_0)^2}{2}\,d\zeta.$$

We conclude that for z in a neighborhood of z_0, we have

$$f(z) = f(z_0) + f'(z_0)(z - z_0) + \frac{f''(z_0)}{2}(z - z_0)^2 + E_f(z, z_0), \tag{81.19}$$

where

$$|E_f(z, z_0)| \leq K \int_\Gamma \frac{|\zeta - z_0|^2}{2}\,|d\zeta| = \frac{K}{6}|z - z_0|^3,$$

and we assume that $|f^{(3)}(\zeta)| \leq K$ for $\zeta \in \Gamma$. More generally, we have the following Taylor's formula:

Theorem 81.3 *If $f : \Omega \to \mathbb{C}$ is analytic in Ω with $|f^{(n+1)}(z)| \leq K$ for $z \in \Omega$, then we have for $z, z_0 \in \Omega$ (with the straight line connecting z and z_0 contained in Ω):*

$$f(z) = f(z_0) + f'(z_0)(z - z_0) + ... + \frac{f^{(n)}(z_0)}{n!}(z - z_0)^n + R_n(z, z_0), \quad (81.20)$$

where $|R_n(z, z_0)| \leq \frac{K}{(n+1)!}|z - z_0|^{n+1}$.

81.25 Cauchy's theorem

We shall now prove that if $f(z)$ is analytic in Ω and Γ is a simple closed curve in Ω enclosing a domain Ω_Γ contained in Ω, then

$$\int_\Gamma f(z)\, dz = 0.$$

To see this we write

$$\int_\Gamma f(z)\, dz = \int_a^b (u(x(t), v(t)) + iv(x(t), y(t)))(\dot{x}(t) + i\dot{y}(t))\, dt,$$

where $\gamma(t) = (x(t), y(t))$ with $a \leq t \leq b$ is a parametrization of Γ. Taking the real part, we get

$$\begin{aligned} \mathrm{Re}(\int_C f(z)\, dz) &= \int_a^b (u(x(t), y(t))\dot{x}(t) - v(x(t), y(t))\dot{y}(t)\, dt \\ &= \int_a^b (u(x, y), -v(x, y)) \cdot (\dot{x}, \dot{y})\, dt. \end{aligned}$$

By the Cauchy-Riemann equations, we have

$$\nabla \times (u, -v) = \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} = 0 \quad \text{in } \Omega_\Gamma,$$

which, proves recalling Stokes' theorem (57.13), that

$$\begin{aligned} \int_a^b (u(x, y), -v(x, y)) \cdot (\dot{x}, \dot{y})\, dt &= \int_\Gamma (u(x, y), -v(x, y)) \cdot ds \\ &= \int_{\Omega_\Gamma} \nabla \times (u, -v) dx dy = 0. \end{aligned}$$

We conclude that $\mathrm{Re}(\int_\Gamma f(z)\, dz) = 0$, and similarly we see that $\mathrm{Im}(\int_\Gamma f(z)\, dz) = 0$ and we have thus proved *Cauchy's theorem*:

Theorem 81.4 (Cauchy's theorem) *If $f(z)$ is analytic in Ω and Γ is a simple closed curve in Ω enclosing a domain contained in Ω, then*

$$\int_\Gamma f(z)\,dz = 0.$$

Note that Γ is not allowed to enclose "holes" of Ω where $f(z)$ is not analytic. For example, we saw above that $\int_\Gamma \frac{1}{z}\,dz = 2\pi i \neq 0$, where Γ is a circle around the origin. This is because Γ encloses the point $z = 0$ where $\frac{1}{z}$ is not analytic.

81.26 Cauchy's representation formula

We prove that if $f(z)$ is analytic in an open domain Ω, and Γ is a simple closed curve in Ω oriented counter-clockwise and bounding the open domain Ω_Γ contained in Ω, then for $z_0 \in \Omega_\Gamma$,

$$f(z_0) = \frac{1}{2\pi i}\int_\Gamma \frac{f(z)}{z - z_0}\,dz, \tag{81.21}$$

which is *Cauchy's representation formula.* Note the counter-clockwise orientation. Further, note that z_0 is not allowed to lie on the curve Γ; we assume that z_0 lies *inside* Γ. Cauchy's formula (81.21) shows that the values of $f(z)$ on Γ alone, determine the values of $f(z)$ in all of Ω_Γ. This shows that an analytic function is not allowed to bring surprises: if we know $f(z)$ on Γ, then we know $f(z)$ in the whole domain Ω_Γ bounded by Γ. The proof follows from realizing that the function

$$g(z) = \frac{f(z) - f(z_0)}{z - z_0} \quad \text{for } z \neq z_0, \quad g(z_0) = f'(z_0),$$

is analytic in Ω, because $g(z)$ is clearly differentiable if $z \neq z_0$ and using a Taylor expansion of $f(z)$, it follows that $g(z)$ is differentiable also at $z = z_0$ with derivative $g'(z_0) = \frac{f''(z_0)}{2}$. Indeed, recalling (81.19) we have

$$g(z) - g(z_0) = \frac{f(z) - f(z_0) - f'(z_0)(z - z_0)}{(z - z_0)} = \frac{f''(z_0)}{2}(z - z_0) + E_f(z, z_0)$$

with $|E_f(z, z_0)| \leq \frac{K}{6}|z - z_0|^2$ and K a bound for $|f^{(3)}(z)|$, which proves the desired result. We conclude that

$$\int_\Gamma \frac{f(z) - f(z_0)}{z - z_0}dz = 0,$$

and using that

$$\int_\Gamma \frac{f(z_0)}{z - z_0}dz = f(z_0)\int_\Gamma \frac{1}{z - z_0}dz = 2\pi i\, f(z_0),$$

we obtain the desired result (81.21). We summarize:

Theorem 81.5 (Cauchy's representation formula) *If $f(z)$ is analytic in an open domain Ω, and Γ is a simple closed curve in Ω oriented counter-clockwise and enclosing the open domain Ω_Γ contained in Ω, then for $z_0 \in \Omega_\Gamma$,*

$$f(z_0) = \frac{1}{2\pi i}\int_\Gamma \frac{f(z)}{z - z_0}\, dz. \tag{81.22}$$

Differentiating with respect to z_0 we obtain the following generalized representation formula:

Theorem 81.6 (Cauchy's generalized representation formula) *If $f(z)$ is analytic in an open domain Ω, and Γ is a simple closed curve in Ω oriented counter-clockwise and enclosing an open domain Ω_Γ in Ω, then for $z_0 \in \Omega_\Gamma$ and $n = 0, 1, 2, ...$,*

$$f^{(n)}(z_0) = \frac{n!}{2\pi i}\int_\Gamma \frac{f(z)}{(z - z_0)^{n+1}}\, dz. \tag{81.23}$$

We note that if z_0 lies outside the region bounded by Γ, then

$$\frac{1}{2\pi i}\int_\Gamma \frac{f(z)}{z - z_0}\, dz = 0,$$

simply because $\int_\Gamma \frac{1}{z-z_0}\, dz = 0$ in this case as a consequence of the fact that $\frac{1}{z-z_0}$ is analytic in a domain containing Γ. Choosing $z_0 \in \Gamma$ leads to a divergent integral beacause of the singularity of the factor $\frac{1}{z-z_0}$, and to define a proper value of the integral in this case leads to the so called *Cauchy principal value*, which we discuss below.

81.27 Taylor's formula: second shot

By using Cauchy's formula we now give another version of Taylor's formula for a function $f(z)$ which is analytic in a neighborhood Ω of a point $z_0 \in \mathbb{C}$. We start writing Cauchy's formula in the form

$$f(z) = \frac{1}{2\pi i}\int_\Gamma \frac{f(\zeta)}{\zeta - z}\, d\zeta, \tag{81.24}$$

where for definiteness we choose Γ to be a counter-clockwise oriented circle around z_0 of radius r contained in Ω. Using the identity

$$\frac{1}{1-q} = 1 + q + q^2 + ... + q^n + \frac{q^{n+1}}{1-q},$$

where $q \in \mathbb{C}$ satisfies $|q| < 1$, setting $q = \frac{z-z_0}{\zeta - z_0}$ with $z \in \Omega$ and $\zeta \in \Gamma$, we can write

$$\frac{1}{\zeta - z} = \frac{1}{\zeta - z_0}\big[1 + \frac{z-z_0}{\zeta - z_0} + ... + (\frac{z-z_0}{\zeta - z_0})^n\big] + \frac{1}{\zeta - z}(\frac{z-z_0}{\zeta - z_0})^{n+1},$$

where we used that

$$\frac{1}{\zeta - z} = \frac{1}{\zeta - z_0 - (z - z_0)} = \frac{1}{\zeta - z_0}\,\frac{1}{1-q}.$$

Insertion into (81.24) now gives

$$f(z) = \frac{1}{2\pi i}\int_\Gamma \frac{f(\zeta)}{\zeta - z_0}\,d\zeta + \frac{z - z_0}{2\pi i}\int_\Gamma \frac{f(\zeta)}{(\zeta - z_0)^2}\,d\zeta + \ldots$$
$$+\frac{(z-z_0)^n}{2\pi i}\int_\Gamma \frac{f(\zeta)}{(\zeta - z_0)^{n+1}}\,d\zeta + R_n(z),$$

where

$$R_n(z) = \frac{(z-z_0)^{n+1}}{2\pi i}\int_\Gamma \frac{f(\zeta)}{(\zeta - z_0)^{n+1}(\zeta - z)}\,d\zeta. \tag{81.25}$$

Using Cauchy's representation formulas we thus obtain the following Taylor formula:

Theorem 81.7 *If $f(z)$ is analytic in a neigborhood Ω of a $z_0 \in \mathbb{C}$, then*

$$f(z) = f(z_0) + f'(z_0)(z - z_0) + \ldots. + \frac{f^{(n)}(z_0)}{n!}(z - z_0)^n + R_n(z), \tag{81.26}$$

where the remainder $R_n(z)$ is given by (81.25) *with Γ a circle around z_0.*

If $\lim_{n\to\infty} R_n(z) = 0$ for z in a neighborhood Ω of z_0, then we obtain the following *power series representation* of $f(z)$ for $z \in \Omega$:

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{(n)}(z_0)}{n!}(z - z_0)^n. \tag{81.27}$$

We conclude by proving that indeed $\lim_{n\to\infty} R_n(z) = 0$ for z in a neighborhood of z_0. We then assume that the disc $D_r(z_0) = \{z \in \mathbb{C} : |z - z_0| \le r$ is contained in the domain Ω of analyticity of $f(z)$ and we assume that $|f(z)| \le M$ for $z \in D_r(z_0)$. Assuming that $|z - z_0| < \frac{r}{2}$, we obtain by inserting absolute values in (81.25) using that $|\zeta - z| \ge \frac{r}{2}$ and $L(\Gamma) = 2\pi r$:

$$|R_n(z)| \le (\frac{|z - z_0|}{r})^{n+1} 2M,$$

which proves that $\lim_{n\to\infty} R_n(z) = 0$ for $|z - z_0| \le \frac{r}{2}$. We can extend the argument to z satisfying $|z - z_0| < r$, and we summarize as follows:

Theorem 81.8 (Taylor's formula) *If $f : \Omega \to \mathbb{C}$ is analytic and $D_r(z_0) = \{z \in \mathbb{C} : |z - z_0| \le r\}$ is contained in Ω and f is bounded on $D_r(z_0)$, then $f(z)$ can be represented as the convergent power series* (81.27) *for $|z - z_0| < r$.*

Power series representations of analytic functions of the form (81.27) play an important role and we devote the next section to this topic starting with the case $z_0 = 0$.

81.28 Power series representation of analytic functions

Consider a series of the form

$$\sum_{m=0}^{\infty} a_m z^m \tag{81.28}$$

where the coefficients $a_m \in \mathbb{C}$ and we assume $z \in \mathbb{C}$. The concepts of convergence and absolute onvergence for (81.28) are direct analogs of the corresponding concepts for series with a_m and z being real, see Chapter *Series*. In particular we say that $\sum_{m=0}^{\infty} a_m z^m$ is absolutely convergent if $\sum_{m=0}^{\infty} |a_m z^m|$ is convergent, and note that an absolutely convergent series is convergent.

Each term of the series (81.28) is analytic in $\mathbb{C}$ and each partial sum

$$\sum_{m=0}^{n} a_m z^m$$

is thus analytic in $\mathbb{C}$. Suppose now that the series (81.28) is convergent for a particular $z = \hat{z}$ with $|\hat{z}| = r$. Since the terms b_m of a convergent series $\sum_{m=0}^{\infty} b_m$ must tend to zero, there is a constant C such that

$$|a_n \hat{z}^n| = |a_n| r^n \leq C \quad n = 0, 1, 2, ...$$

Suppose now that $|z| < r$. We then have

$$\sum_{n=0}^{\infty} |a_n z^n| = \sum_{n} |a_n r^n| |\frac{z}{r}|^n \leq C \sum_{n=0}^{\infty} |\frac{z}{r}|^n < \infty,$$

because $|\frac{z}{r}| < 1$. This proves that $\sum_{n=0}^{\infty} a_n z^n$ is absolutely convergent for $|z| < r$ and is thus convergent for $|z| < r$.

We say that the *radius of convergence* of $\sum_{n=0}^{\infty} a_n z^n$ is equal to r, if $\sum_{n=0}^{\infty} a_n z^n$ is convergent for $|z| < r$ but not convergent for some z with $|z| \geq r$.

One can (easily) show that inside its radius of convergence r a series $\sum_{n=0}^{\infty} a_n z^n$ is differentiable with

$$(\sum_{n=0}^{\infty} a_n z^n)' = \sum_{n=1}^{\infty} n a_n z^{n-1},$$

where the termwise differentiated series $\sum_{n=1}^{\infty} n a_n z^{n-1}$ is also convergent for $|z| < r$.

More generally, we consider power series of the form

$$\sum_{m=0}^{\infty} a_m (z - z_0)^m, \tag{81.29}$$

where we made a shift of variable from z to $z - z_0$ with $z_0 \in \mathbb{C}$ given. The notion of convergence and raduis of convergence extend in the obvious way. Of course, (81.29) connects to the Taylor series of $f(z)$ at z_0 with $a_m = \frac{f^{(m)}(z_0)}{m!}$.

EXAMPLE 81.13. The series

$$\sum_{n=0}^{\infty} \frac{z^n}{n!}$$

is convergent for any fixed $z \in \mathbb{C}$, since $n! = 1 \cdot 2 \cdot 3 \cdots n$ grows much quicker than r^n for any $r > 0$. We can thus differentiate termwise and we get

$$(\sum_{n=0}^{\infty} \frac{z^n}{n!})' = \sum_{n=1}^{\infty} \frac{z^{n-1}}{(n-1)!} = \sum_{n=0}^{\infty} \frac{z^n}{n!},$$

which shows that $\sum_{n=0}^{\infty} \frac{z^n}{n!}$ satisfies the differential equation $u'(z) = u(z)$ with the initial condition $u(0) = 1$. It follows that

$$\exp(z) = \sum_{n=0}^{\infty} \frac{z^n}{n!}. \tag{81.30}$$

Alternatively, this follows by noting that this is the Taylor series representation of $f(z) = \exp(z)$ around $z_0 = 0$, noting that $f^{(n)}(z) = \exp(z)$ for $n = 1, 2, ...,$.

Using that $\cos(z) = \frac{1}{2}(\exp(iz) + \exp(-iz))$ and $\sin(z) = \frac{1}{2i}(\exp(iz) - \exp(-iz))$, we obtain the following Taylor series representations valid for $z \in \mathbb{C}$:

$$\cos(z) = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!}, \qquad \sin(z) = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)!}.$$

EXAMPLE 81.14. Another basic example is given by

$$\log(1+z) = \sum_{n=1} \frac{(-1)^{n-1}}{n} z^n \quad \text{for } |z| < 1,$$

which is readily obtained differentiating $\log(1+z)$.

81.29 Laurent series

Consider a series of the form

$$\sum_{m=1}^{\infty} b_m z^{-m}, \tag{81.31}$$

obtained by replacing z by $\frac{1}{z}$ in a power series $\sum_{m=1}^{\infty} b_m z^m$ with radius of convergence r. The series (81.31) will thus converge for $|z| > r$. More generally we may consider a *Laurent series* of the form

$$f(z) = \sum_{m=0}^{\infty} a_m z^m + \sum_{m=1}^{\infty} b_m z^{-m}, \tag{81.32}$$

which we assume to be convergent in an annulus $\{r_1 < |z| < r_2\}$. The function $f(z)$ defined by (81.32) is analytic in the annulus $\{r_1 < |z| < r_2\}$. Conversely, if $f(z)$ is analytic in the annulus $\{r_1 < |z| < r_2\}$, then $f(z)$ admits the Laurent series expansion (81.32) with the coefficients a_m and b_m being given by

$$a_m = \frac{1}{2\pi i}\int_\Gamma \frac{f(\zeta)}{\zeta^{m+1}}\,d\zeta, \quad b_m = \frac{1}{2\pi i}\int_\Gamma f(\zeta)\zeta^{m-1}\,d\zeta, \tag{81.33}$$

where Γ is a simple closed counter-clockwise oriented curve in the annulus encircling the origin. The formual for the coefficients is obtaine by multiplying by a proper power of z and integrating around Γ.

We may generalize to shifts of the origin to a given point z_0 replacing z by $z - z_0$.

Example 81.15. We have

$$\begin{aligned} \frac{1}{1-z} &= \sum_{m=0}^{\infty} z^m \quad \text{for } |z| < 1, \\ \frac{1}{1-z} &= \frac{-1}{z(1-z^{-1})}\sum_{m=1}^{\infty} z^{-m} \quad \text{for } |z| > 1. \end{aligned}$$

Example 81.16.

81.30 Residue Calculus: simple poles

Let $f(z)$ be analytic in a simply connected open domain Ω, except at an isolated point $z_0 \in \Omega$, and let Γ be a simple closed curve in Ω oriented counter-clockwise with z_0 contained in the open domain Ω_Γ bounded by Γ. We say that the simple closed curve Γ *surrounds* z_0 counter clockwise. In general the integral

$$\int_\Gamma f(z)\,dz$$

will then not be zero, but the integral will have the same value for any simple closed curve Γ_1. To see this we introducing the two coinciding curves $\Gamma_3^{\pm}$ with opposite orientation joining Γ and Γ_1 according to Fig. **??**, and by joining the curves Γ, Γ_3^+, Γ_1 and Γ_3^- we obtain a single closed curve enclosing a domain where $f(z)$ is analytic (that is, not containing z_0 in its interior) over which the integral of $f(z)$ vanishes becasue of Cauchy's theorem. Thus, noting that the intergrals over Γ_3^+ and Γ_3^- cancel, we have

$$0 = \int_\Gamma f(z)\,dz + \int_{\Gamma_2} f(z)\,dz = \int_\Gamma f(z)\,dz - \int_{\Gamma_1} f(z)\,dz$$

where we used that the orientation of Γ_1 and Γ_2 are reversed. It follows that the integral over Γ is equal to the integral ovber Γ_1.

A fig of Γ, Γ_3^+, Γ_1 and Γ_3^- and z_0

Suppose now that $f(z)$ has the form

$$f(z) = \frac{g(z)}{z - z_0},$$

where $g(z)$ is analytic in Ω and $z_0 \in \Omega$. We then say that $f(z)$ has a *simple pole* at $z = z_0$. We have by Cauchy's representation formula with Γ a simple closed curve surrounding z_0 counter-clockwise,

$$\int_\Gamma f(z)\,dz = \int_\Gamma \frac{g(z)}{z - z_0}\,dz = 2\pi i g(z_0).$$

The value $g(z_0)$ is called the *residue* of $f(z)$ at z_0, which we denote by $\operatorname{Res} f(z_0)$, and thus

$$Res\, f(z_0) = g(z_0) = \lim_{z\to z_0} (z - z_0) f(z).$$

Example 81.17. Let $f(z) = \frac{z}{z-1}$ and let Γ be the circe $\gamma(t) = (\cos(t) - 1, \sin(t))$ with $0 \le t \le 2\pi$, surrounding $(1, 0)$ counter-clockwise. By the Residue Theorem, we have since obviously $Res\, f(1) = 1$

$$\int_\Gamma \frac{z}{z - 1}\,dz = 2\pi i.$$

Example 81.18. To evaluate $\int_\Gamma f(z)\,dz$, where $f(z) = \frac{1}{e^z - 1}$ and Γ is a circle centered at the origin and oriented counter-clockwise, we note that

$$f(z) = \frac{z}{e^z - 1}\frac{1}{z} = \frac{g(z)}{z}$$

with

$$\frac{1}{g(z)} = \frac{e^z - 1}{z} = h(z).$$

Since $\lim_{z\to 0} h(z) = 1$, we have $\operatorname{Res}\ f(0) = g(0) = 1$, and thus $\int_\Gamma f(z)\,dz = 2\pi i$.

81.31 Residue Calculus: poles of any order

Suppose now $f(z)$ has a (multiple) *pole of order* $n = 2, 3, ...,$ at z_0, that is, $f(z)$ is of the form

$$f(z) = \frac{g(z)}{(z - z_0)^n},$$

with $g(z)$ analytic in a neighborhood of z_0. By Cauchy's generalized representation formula we have if Γ is a simple closed curved surrounding z_0 counter-clockwise:

$$\int_\Gamma f(z)\,dz = \int_\Gamma \frac{g(z)}{(z - z_0)^n}\,dz = \frac{2\pi i}{(n-1)!}\,g^{(n-1)}(z_0).$$

We now extend the definition of the residue $\mathrm{Res}\,f(z_0)$ to a pole of order $n = 1, 2, ...,$ by setting

$$\mathrm{Res}\,f(z_0) = \frac{g^{(n-1)}(z_0)}{(n-1)!},$$

and thus we have again

$$\int_\Gamma f(z)\,dz = 2\pi i\,\mathrm{Res}\,f(z_0).$$

Example 81.19. The function

$$f(z) = \frac{1}{(z-1)^2(z-3)}$$

has a pole of order 2 at $z = 1$ and order 1 at $z = 3$. We compute $Res\,f(3) = \frac{1}{4}$, and further $Res\,f(1) = -\frac{1}{4}$ since $\frac{d}{dz}\frac{1}{z-3} = -\frac{1}{(z-3)^2} = -\frac{1}{4}$ if $z = 1$.

81.32 The Residue Theorem

We now prove the following basic result of residue calculus:

Theorem 81.9 (The Residue Theorem) *Let $f(z)$ be analytic in a simply connected open domain Ω, except at finitely many isolated points z_1, z_1,...,z_n in Ω where $f(z)$ has simple or multiple poles, and let Γ be a simple closed curve in Ω surrounding all the z_m counter-clockwise. Then*

$$\int_\Gamma f(z)\,dz = \sum_{m=1}^{n} 2\pi i Res\,f(z_m).$$

The result folllows by surrounding each of the z_m with a little circle $= \Gamma_m$ inside Γ oriented clockwise. By Cauchy's theorem we then have

$$\int_\Gamma f(z)\, dz + \sum_{m=1}^{n} \int_{\Gamma_m} f(z)\, dz = 0,$$

arguing as in the Section on Residue Calculus: simple poles, from which follows that

$$\int_\Gamma f(z)\, dz = -\sum_{m=1}^{n} \int_{\Gamma_m} f(z)\, dz = 2\pi i \sum_{m=1}^{n} \operatorname{Res} f(z_m),$$

which proves the desired result.

Example 81.20. We compute

$$I = \int_\Gamma \frac{4-3z}{z^2-z}\, dz = \int_\Gamma \frac{4-3z}{z(z-1)}\, dz$$

where Γ is a simple closed curve surrounding counter-clockwise the two simple poles $z = 1$ and $z = 0$ of $\frac{4-3z}{z^2-z}$ and get $I = 2\pi i(-4+1) = -6\pi i$.

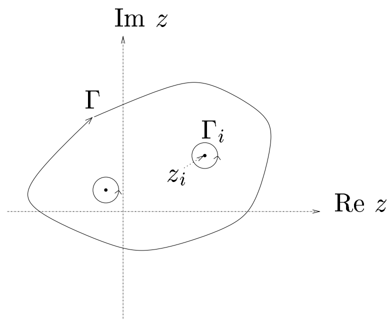


FIGURE 81.11. A curve Γ and curves Γ_i, surrounding poles of $f(z)$.

81.33 Computation of $\int_0^{2\pi} R(\sin(t), \cos(t))\, dt$

Consider an integral of the form

$$\int_0^{2\pi} R(\sin(t), \cos(t))\, dt$$

where $R(x,y)$ is a rational function of $x, y \in \mathbb{R}$. By the substitution

$$z = e^{it}, \quad dz = ie^{it}dt = iz\,dt$$
$$\cos(t) = \frac{e^{it}+e^{-it}}{2} = \frac{1}{2}(z+\frac{1}{z})$$
$$i\sin(t) = \frac{e^{it}-e^{-it}}{2} = \frac{1}{2}(z-\frac{1}{z}),$$

the integral is converted into

$$\int_{|z|=1} R(\frac{z^2-1}{2iz}, \frac{z^2+1}{2z})\frac{dz}{iz},$$

which can be evaluated using residue calculus, provided the integrand has no poles on $|z| = 1$.

Example 81.21. We compute

$$I = \int_0^{2\pi} \frac{dt}{a+\cos(t)}\,dt,$$

where $a > 1$ is a constant. Using the transformation just indiciated we get

$$I = -2i\int_{|z|=1} \frac{dz}{z^2+2az+1} = -2i\int_{|z|=1} \frac{dz}{(z-\alpha)(z-\beta)},$$

where $\alpha = -a + \sqrt{a^2-1}$ and $\beta = -a - \sqrt{a^2-1}$. Since $|\alpha| < 1$ and $|\beta| > 1$, the residue at α is $\frac{1}{\alpha-\beta}$ and thus $I = \frac{2\pi}{\sqrt{a^2-1}}$.

81.34 Computation of $\int_{-\infty}^{\infty} \frac{p(x)}{q(x)}\,dx$

Integrals of the form

$$\int_{-\infty}^{\infty} f(x)\,dx \tag{81.34}$$

can be evaluated using residue calculus under the assumption that the extended function $f(z)$ with $z \in \mathbb{C}$ has no poles on the real axis and that $|f(z)| \leq M|z|^{-2}$ for $|z|$ large. We start out showing how to use residue calculus to compute the integral

$$I = \int_{-\infty}^{\infty} \frac{1}{1+x^2}\,dx,$$

(thus without using that $\arctan(x)$ is a primitive function of $\frac{1}{1+x^2}$). We write

$$I = \lim_{R\to\infty}\int_{-R}^{R} \frac{1}{1+x^2}\,dx = \lim_{R\to\infty}\int_{\Gamma_R} f(z)\,dz,$$

where $f(z) = \frac{1}{1+z^2}$ and Γ_R is the boundary of the semi-disc $|z| \leq R$ with Re $z = x \geq 0$. This follows from the fact that

$$\lim_{R\to\infty} \int_{\Gamma_R^+} f(z)\, dz = 0,$$

where Γ_R^+ is the upper part of the semi-circle with Re $z = x > 0$. By the Residue theorem we have

$$\int_{\Gamma_R} f(z)\, dz = 2\pi i \frac{1}{2i} = \pi$$

since the residue of $f(z) = \frac{1}{(z-i)(z+i)}$ inside Γ_R is equal to $\frac{1}{z+i}$ with $z = i$. We conclude that

$$\int_{-\infty}^{\infty} f(x)\, dx = \pi,$$

which of course conforms with the result obtained using that $\frac{d}{dx}\arctan(x) = \frac{1}{1+x^2}$.

The same technique can be used if $f(x) = \frac{p(x)}{q(x)}$ is a rational function with the the degree of $q(x)$ two units higher than that of the numerator $p(x)$. The same technique can be used to evaluate the Fourier transform (cf below) of $\frac{p(x)}{q(x)}$:

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{p(x)e^{i\xi x}}{q(x)}\, dx.$$

81.35 Applications to potential theory in $\mathbb{R}^2$

There is a strong coupling between analytic functions and potential theory in $\mathbb{R}^2$, because if $f(z) = u(x,y) + iv(x,y)$ is analytic in Ω, then the real and imaginary parts $u(x,y)$ and $v(x,y)$ are harmonic in Ω, that is, $\Delta u = \Delta v = 0$ in Ω. Conversely, as we saw above, if $u(x,y)$ is harmonic in a simply connected domain Ω in $\mathbb{R}^2$, then there exists a harmonic conjugate function $v(x,y)$ uniquely determined up to a constant such that $f(z) = u(x,y) + iv(x,y)$ is analytic in Ω, see Problem 81.10. The Cauchy-Riemann equations state that $\nabla u = \nabla \times v$:

$$\nabla u = (\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}) = \nabla \times v = (\frac{\partial v}{\partial y}, -\frac{\partial v}{\partial x}).$$

from which follows that ∇u and ∇v are orthogonal:

$$\nabla u \cdot \nabla v = \frac{\partial u}{\partial x}\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\frac{\partial v}{\partial y} = \frac{\partial v}{\partial y}\frac{\partial v}{\partial x} - \frac{\partial v}{\partial x}\frac{\partial v}{\partial y} = 0.$$

We conclude that the level curves of $u(x,y)$ and its conjugate $v(x,y)$ are orthogonal. We note that level curves of $u(x,y)$ and $v(x,y)$ in the $z=(x,y)$-plane correspond to the level lines $u=$ constant and $v=$ constant of the analytic function $w=u+iv$ in the $w=(u,v)$-plane.

In fact, much of the interest in analytic functions comes from the connection to potential theory in $\mathbb{R}^2$. Today, computational methods capable of solving also problems in $\mathbb{R}^3$, have changed this picture and analytic functions now play a less important role in areas of applications such as fluid and solid mechanics.

Applications to fluid mechanics typically concern incompressible irrotational flow in 2d with $u(x,y)$ representing a velocity potential and $v(x,y)$ an associated so called *stream function*. The velocity U of the flow is then given by $U=\nabla u=\nabla\times v$

$$U=\nabla u=(\frac{\partial u}{\partial x},\frac{\partial u}{\partial y})=\nabla\times v=(\frac{\partial v}{\partial y},-\frac{\partial v}{\partial x}).$$

We have $\nabla\cdot U=\Delta u=0$ and $\nabla\times U=-\Delta v=0$ and thus U is incompressible and irrotational. The level curves of $u(x,y)$ with normal ∇u correspond to equi-potential curves, and the level curves of v with normal $\nabla v=-\nabla\times u$ will correspond to the *streamlines* followed by a fluid particle moving with the velocity U. We conclude that each analytic function $f(z)=u(x,y)+iv(x,y)$ may be associated to a particular stationary incompressible and irrotational fluid flow, and the level curves of u and v form a mutually orthogonal set of curves with the level curves of v describing the streamlines of the flow.

In applications to electromagnetics, $u(x,y)$ represents an electric potential with ∇u an electric field, and the level curves of $v(x,y)$ represent the curves traced by electrically charged particles in the electric field.

In applications to heat flow $u(x,y)$ may represent temperature and the level curves for u thus become isolines for temperature and ∇u is proportional to the heat flow.

EXAMPLE 81.22. **(Flow in a corner)** The function $w=u+iv=z^2$ describes a certain flow in the quarter-plane $\{z=x+iy:x,y\geq 0\}$, with corresponding potential $u(x,y)=x^2-y^2$ and stream-function $v(x,y)=2xy$, *see* Fig. **??** The equi-potential lines $u(x,y)=$ constant and streamlines $v(x,y)=$ constant in the (x,y)-plane are the images of the lines $u=$ constant and $v=$ constant under the mapping $z=w^{1/2}$ from the halfplane $\{w=u+iv:v\geq 0\}$ onto the quarter-plane $\{z=x+iy:x,y\geq 0\}$.

EXAMPLE 81.23. **(The spinning tennis ball)** We consider two types of rotation-free incompressible flow around the unit disc in two dimensions. The first flow is given by the function $f(z)=z+\frac{1}{z}$, which in polar coordinates with $z = re^{i\theta}$ takes the form

$$f(z) = u(r,\theta) + iv(r,\theta) = (r + \frac{1}{r})\cos(\theta) + i(r - \frac{1}{r})\sin(\theta). \quad (81.35)$$

This represents a symmetric flow with the velocity equal to $(1,0)$ away from the disc, and the level curves of $v(r,\theta)$ give the streamlines of the flow around the disc, see Fig. **??**. The second flow is a flow circulating around the disc given by

$$g(r,\theta) = -\frac{iK}{2\pi}\log(z) = \frac{K}{2\pi}\theta + i(-\frac{K}{2\pi}\log(r)) \quad (81.36)$$

Consider now the flow given by $f(z) + g(z)$ with stream-function $(r - \frac{1}{r})\sin(\theta) - \frac{K}{2\pi}\log(r)$. We may consider this to be the flow around a spinning tennis ball in a horisontal stream of air, see Fig. **??**

We now recall *Bernouilli's law* stating that for steady inviscid irrotational flow, the quantity

$$p + \frac{|U|^2}{2}$$

is constant because $\nabla(p + \frac{|U|^2}{2}) = 0$, which follows by direct computation, see Problem 81.13. We conclude that high velocity implies low pressure. Now inspecting Fig. **??** we see that the velocity is high below the ball (dense level curves of the stream function), and thus the pressure is low below the ball and thus there will be a resulting force downward, which is referred to as *lift*. This is the reason a top-spin in tennis is so efficient in bringing down the ball inside the lines. The more top spin the more curved path of the ball! Björn Borg was one of the first to really exploit this law of mechanics. One can show that the *lift* is proportional to the *circulation* given by

$$\int_\Gamma u \cdot ds,$$

where Γ is the unit circle oriented counter-clockwise. The circulation of the flow given by (81.35) is equal to zero because of symmetry, while the circulation of the flow given by (81.36) is equal to K. The lift of the spinning tennis ball gives a hint to the mechanism behind flying. In fact, the design of an airplane wing with non-symmetric cross section and a sharp trailing edge creates a circulation around the wing which causes a lift, see Fig. **??**. Note the counter-clockwise vortex behind the wing which is balanced by a clockwise circulation around the wing, because the total rotation of the flow must be zero.

EXAMPLE 81.24. (**Flow through an aperture**) The function

$$\begin{aligned} z = \sin(w) &= \frac{1}{2i}e^{i(u+iv)} - e^{-i(u+iv)} \\ &= \sin(u)\cosh(v) + i(\cos(u)\sinh(v)), \end{aligned}$$

maps the strip $\{w = u + iv : \frac{-\pi}{2} < u < \frac{\pi}{2}, v \in \mathbb{R}\}$ onto $\{z = x + iy : y \neq 0 \text{ if } |x| > 1\}$, that is, the whole plane minus the two half-lines $\{x + iy : |x| > 1, y = 0\}$. The corresponding inverse function $w = f(z) = \arcsin(z) = \sin^{-1}(z)$ may be viewed as the potential for flow through an aperture, see Fig. **??**. The streamlines are hyperbolas and the equipotential lines are ellipses.

EXAMPLE 81.25. (**Discontinuous electric potential**) Consider the function $f(z) = u(x, y) + iv(x, y) = i\log(z) = i\log(|z|) - \operatorname{Arg} z$ in the right half-plane $\{z \in \mathbb{C} : \operatorname{Re} z \geq 0\}$. We have $u(x, y) = \arctan(\frac{y}{x})$ and $v(x, y) = \log(r)$ with $r = (x^2 + y^2)^{1/2}$ and we plot the equi-potential and level curves of the curves in Fig. **??**. Note that the potential $u(x, y)$ approaches the value $\frac{\pi}{2}$ for x tending to zero if $y > 0$ and the value $-\frac{\pi}{2}$ for x tending to zero if $y < 0$, corresponding to discontinuous boundary values for $x = 0$.

Chapter 81 Problems

81.1. (a) Prove that $f(z) = e^z$ is analytic and that $f'(z) = e^z$. (b) Prove that $\sin(z)$ and $\cos(z)$ are analytic with derivatives $\cos(z)$ and $-\sin(z)$, respectively.

81.2. It is possible to view an analytic function $f : \mathbb{C} \to \mathbb{C}$ as a function $F : \mathbb{R}^2 \to \mathbb{R}^2$ if we set $f(z) = u(x, y) + iv(x, y)$, $z = x + iy$ and $F(x, y) = (u(x, y), v(x, y))$. Explain the connection between the Jacobian F' of $F(x, y)$ and the derivative f', and motivate the Cauchy-Riemann equations this way.

81.3. What happens if we try to choose $z_0 \in \Gamma$ in Cauchy's representation formula?

81.4. Prove Liouville's theorem stating that if $f(z)$ is analytic in the whole complex plane and bounded, then $f(z)$ is constant. Hint: Use the representation formula for $f'(z)$ with Γ a circle with large radius.

81.5. Prove Morera's theorem stating that if $f : \Omega \to \mathbb{C}$ satisfies $\int : \Gamma f(z)\, dz = 0$ for all sinmple closed curves in Ω, then $f(z)$ is analytic in Ω. Hint: Define $F(z) = \int_{\Gamma_z} f(\zeta)\, d\zeta$, where Γ_z is a courve joining a fixed point z_0 with the variable point $z \in \Omega$. Show independence of the specific choice of Γ_z and then that $F'(z) = f(z)$.

81.6. Prove that a Möbius transformation maps every straight line or circle in the complex plane into a circle or straight line. Hint: write $w = \frac{az+b}{cz+d}$ in the form $w = -\frac{ad-bc}{c}\frac{1}{cz+d} + \frac{a}{c}$.

81.7. Compute (a) $\int_0^{2\pi} \frac{d\theta}{5-3\sin(\theta)}$, (b) $\int_{-\infty}^{\infty} \frac{x}{1+x^4}\, dx$.

81.8. Prove that $\int_{-\infty}^{\infty} \frac{\sin\theta}{\theta} = 2\pi$.

81.9. Prove (81.10).

81.10. Prove that if $u(x, y)$ is harmonic in a simply connected domain Ω, then there exists a function v such that $u + iv$ satisfies the Cauchy-Riemann equations in Ω. Hint use the central result of Chapter *Potential fields.*

81.11. Construct your own examples of 2d irrotational potential flow, electrostatics, and heat flow, by combining elementary functions such as z^α, e^z, $\log(z)$, $\sin(z)$, $\sinh(z)$ and Möbius transformations.

81.12. Give a different proof of Cauchy's representation theorem using that $\frac{g(z)}{z-z_0}$ is a analytic in the domain $\Omega_\epsilon = \{z \in \Omega : |z - z_0| > \epsilon\}$, so that $\int_{\Gamma_\epsilon} \frac{g(z)}{z-z_0}\, dz = 0$, where Γ_ϵ is the boundary of Ω_ϵ. Then let ϵ tend to zero.

81.13. Show that if the fluid velocity $u = (u_1, u_2)$ defined in a domain Ω in $\mathbb{R}^2$ satisfies $\nabla \cdot u = \nabla \times u = 0$ and solves the stationary momentum equation $(u \cdot \nabla)u + \nabla p - \Delta u = 0$ in Ω, then $\nabla(p + \frac{|u|^2}{2}) = 0$ in Ω. This proves Bernouilli's Law stating that $p + \frac{|u|^2}{2}$ is constant so that high velocity corresponds to low pressure.

82
Fourier series

> Yesterday was my 21st birthday, at that age Newton and Pascal had already acquired many claims to immortality. (Fourier 1787, age 21)

82.1 Introduction

We give in the following two chapters a short account of *Fourier analysis* starting with *Fourier series* in this chapter and continuing in the next chapter to *Fourier transforms.* The basic idea is to represent (or approximate) given functions as linear combinations of trigonometric functions. We have met the same general idea in the Chapter Piecewise linear approximation, where we studied approximation of given functions as a linear combination of piecewise polynomials. Fourier representaions have particular properties which are useful in for example signal/image processing with important applications to e.g. computer tomography. In recent years variants of Fourier techniques referred to as *Wavelets* have been developed with applications to for example compression of images. We touch this topic at the end of the Chapter *Fourier transforms.*

Fourier (1768-1830) used trigonometric series in his famous Théorie analytique de chaleur (1822) to study properties of solutions of the heat equation. The idea of expressing a general function as or Fourier series (or as a power series) has influenced the development of mathematical analysis profoundly with the driving force being the formidable success of these techniques for certain classes of problems, for example linear constant coefficient differential equations. However, as any highly specialized tool or

FIGURE 82.1. Fourier

organism, these techniques have not been able to adapt to the needs of a changing world with computational methods for nonlinear differential equations taking over as work-horse in applications. Nevertheless, Fourier analyis still plays a fundamental role for the basic understanding of many phenomena.

We start with Fourier series in complex form and then pass to the real form as a special case. Fourier series concern functions $f : \mathbb{R} \to \mathbb{C}$ which are *periodic* with a certain period $a > 0$, that is $f(x+a) = f(x)$ for $x \in \mathbb{R}$. We often normalize to $a = 2\pi$ and thus consider 2π-periodic functions $f : \mathbb{R} \to \mathbb{C}$ satisfying $f(x + 2\pi) = f(x)$ for $x \in \mathbb{R}$. Usually we restrict attention to real-valued functions $f : \mathbb{R} \to \mathbb{R}$. Fourier transforms concern non-periodic functions $f : \mathbb{R} \to \mathbb{C}$.

We shall see that representing a given 2π-periodic function $f(x)$ as a Fourier series corresponds to expressing $f(x)$ as a linear combination of a certain set a trigonometric functions $\{e_m(x)\}$:

$$f(x) = \sum_m c_m e_m(x) \tag{82.1}$$

with certain coefficients $c_m \in \mathbb{C}$. We thus view the functions $e_m(x)$ as *basis functions* and express a general function $f(x)$ as a certain *linear combination* of basis functions. For example $f(x) = 2\sin(3x) + 7\sin(15x)$ is a linear combination of the two basis functions $sin(3x)$ and $sin(15x)$ with coefficients 3 and 5, see Fig. **??**.

A fig with the linear combination $2\sin(3x) + 7\sin(15x)$

The trigonometric basis functions $e_m(x)$ used in Fourier series are of the form

$$\sin(mx), \quad \cos(mx), \quad m = 0, 1, 2, ..., \quad \text{(real Fourier series)} \tag{82.2}$$

$$e^{imx} = \cos(mx) + i\sin(mx), \quad m = 0, \pm 1, \pm 2, ..., \quad \text{(complex Fourier series).} \tag{82.3}$$

Each basis function or "harmonic" $sin(mx)$, $\cos(mx)$ or e^{imx}, is periodic with period $\frac{2\pi}{|m|}$ and (angular) *frequency* $|m|$. The larger $|m|$ is the higher is the frequency and the quicker do the basis functions $sin(mx)$, $\cos(mx)$ and e^{imx} "oscillate". The series (82.1) expresses $f(x)$ as a linear combination of basis functions of increasing frequencies. Since the basis functions are all periodic with period 2π, so is the linear combination $f(x)$.

The basis functions (82.2) and (82.3) are *orthogonal* with respect to the $L_2(-\pi, \pi)$ scalar product

$$(v, w) = \int_{-\pi}^{\pi} v(x)\overline{w(x)}\, dx \tag{82.4}$$

with $\overline{w(x)}$ the complex conjugate of $w(x)$, with corresponding norm $\|v\| = (v, v)^{1/2}$. The orthogonality makes the coeffcients c_m directly computable upon taking the $L_2(-\pi, \pi)$ scalar product of (82.1) with e_m to give

$$c_m = \frac{(f, e_m)}{(e_m, e_m)}.$$

.

82.2 Warm up I: orthonormal basis in $\mathbb{C}^n$

To prepare we consider an analogous situation in $\mathbb{C}^n$: We recall that $\mathbb{C}^n$ is the set of ordered n-tuples $x = (x_1, ..., x_n)$ with $x_k \in \mathbb{C}$ for $k = 1, ..., n$. The scalar product (x, y) of two vectors x and y in $\mathbb{C}^n$ is defined by $x \cdot y = (x, y) = \sum_{j=1}^n x_j\overline{y_j}$, with corresponding norm $|x| = (x, x)^{1/2}$

Let now $\{g_1, ..., g_n\}$ be a set of n vectors in $\mathbb{C}^n$, that is each $g_k = (g_{k1}, ..., g_{kn})$ is a vector in $\mathbb{C}^n$ with components $g_{kj} \in \mathbb{C}$. We recall that the set $\{g_1, ..., g_n\}$ is an orthonormal basis in $\mathbb{C}^n$ if the g_k are mutually orthogonal and have norm equal to one, that is

$$(g_k, g_m) = 0 \quad \text{if } k \neq m, \quad \text{and } |g_m| = 1 \quad \text{for } m = 1, ...n.$$

If $\{g_1, ..., g_n\}$ is an orthonormal basis, then we can express a given vector $u \in \mathbb{C}^n$ as a linear combination of basis vectors in the form

$$u = \sum_{k=1}^n c_m g_m, \quad \text{where } c_m = (u, g_m) \quad \text{for } m = 1, ...n,$$

where the fact that $c_m = (u, g_m)$ follows by taking the scalar product and using the orthonormality.

82.3 Warm up II: Series

We recall from Chapter *Series* that a series $\sum_{m=1}^{\infty} \alpha_k$ with coefficients $\alpha_k \in \mathbb{C}$, is said to be convergent if the sequence $\{s_n\}_{n=1}^{\infty}$ of partial sums $s_n = \sum_{m=1}^{n} \alpha_k$ converges as n tends to infinity. The series is said to be absolutely convergent if $\sum_{m=1}^{\infty} |\alpha_k|$ is convergent, which is the same as requiring the sequence of partial sums $\hat{s}_n = \sum_{m=1}^{n} |\alpha_k|$ to be bounded above, that is $\hat{s}_n \leq K$ for $n = 1, 2, ...$, where K is a positive constant. For a series with non-negative terms, the concepts of convergence and absolute convergence coincide. A typical example of a positive (absolutely) convergent series is given by $\sum_{m=1}^{\infty} m^{-2}$. To see that $s_n = \sum_{m=1}^{n} m^{-2}$ is bounded above, we use the fact that

$$s_n \leq 1 + \sum_{m=2}^{n} \int_{m-1}^{m} x^{-2}\,dx \leq 1 + \int_{1}^{n} x^{-2}\,dx \leq 1 + [-x^{-1}]_1^n \leq 2.$$

The same argument shows that $\sum_{m=1}^{\infty} m^{-\alpha}$ is convergent if $\alpha > 1$.

We also recall that an alternating series of the form $\sum_{m=1}^{\infty} (-1)^m a_m$ with $\{a_m\}$ a decreasing positive sequence tending to zero, is convergent.

82.4 Complex Fourier series

A series of the form

$$\sum_{m=-\infty}^{\infty} c_m e^{imx} = \sum_{1}^{\infty} c_{-m} e^{-imx} + c_0 + \sum_{1}^{\infty} c_m e^{imx}, \qquad (82.5)$$

where $x \in \mathbb{R}$, is said to be a *Fourier series* with *Fourier coefficients* $c_m \in \mathbb{C}$, $m = 0, \pm 1, \pm 2, ...$ The corresponding *truncated Fourier series*

$$\sum_{m=-n}^{n} c_m e^{imx} = \sum_{m=1}^{n} c_{-m} e^{-imx} + c_0 + \sum_{m=1}^{n} c_m e^{imx}, \qquad (82.6)$$

where $n = 1, 2, ...$, may be viewed as a finite linear combination of the set of basis functions

$$\{1, e^{\pm ix}, e^{\pm i2x}, ..., e^{\pm inx}\}$$

with coefficients c_m.

The orthogonality of the basis functions $\{e^{imx}\}$ is expressed by:

$$\int_{-\pi}^{\pi} e^{imx} e^{-ikx}\,dx = \begin{cases} 0 & \text{if } k \neq m, \\ 2\pi & \text{if } k = m, \end{cases} \qquad (82.7)$$

which follows by direct integration.

We shall typically consider cases with the Fourier coefficients c_m satisfying for some positive constant K,

$$|c_m| \leq Km^{-2}, \quad m = \pm 1, \pm 2, \dots \tag{82.8}$$

In this case the series (82.5) converges absolutely for all x, since

$$\sum_{-\infty}^{\infty} |c_m e^{imx}| = \sum_{-\infty}^{\infty} |c_m| \leq |c_0| + 2K \sum_{m=1}^{\infty} m^{-2} < \infty,$$

and thus defines a function $f : \mathbb{R} \to \mathbb{C}$ represented by a converging Fourier series:

$$f(x) = \sum_{m=-\infty}^{\infty} c_m e^{imx}. \tag{82.9}$$

The series (82.9) gives a *spectral decomposition* of $f(x)$ into harmonics e^{imx} with different amplitudes c_m. The series (82.9) thus gives a description of the function $f(x)$ in terms of amplitudes of different harmonics included in $f(x)$. In musical terms we may think of $f(x)$ as a "chord" built by a number of "tones" $c_m e^{imx}$ of different frequencies m and amplitudes c_m. A spectral decomposition of a "chord" $f(x)$ would display the "tones" building the "chord", try *The Sound of Functions* in the Mathematics Laboratory for an direct experience.

We note that the basis functions $\{e^{imx}\}$ have *global support*, that is, each basis function e^{imx} is nonzero for all $x \in \mathbb{R}$. The basis functions $\{e^{imx}\}$ thus combines the following properties: orthogonality and global support. We contrast this to the 'hat functions' which are the basis functions for continuous piecewise linear approxiamtion: the hat functions have local support but are not orthogonal. The best combination would be orthogonality together with local support. So-called *wavelets* introduced in recent years combine these properties.

Suppose now that $f(x)$ is defined by a converging Fourier series (82.9). Multiplying by e^{-imx} with $m = 0, \pm 1, \pm 2, \dots$ and integrating over the interval $[-\pi, \pi]$, and using the orthogonality properties (82.7), we find that

$$c_m = c_m(f) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{-imx}\, dx, \tag{82.10}$$

where we indicated the dependence of the Fourier coefficient $c_m = c_m(f)$ on the function $f(x)$. We thus have the *Fourier series representation*

$$f(x) = \sum_{m=-\infty}^{\infty} c_m(f) e^{imx}, \tag{82.11}$$

expressing $f(x)$ as a linear combination of different harmonics e^{imx} with different frequencies, where the Fourier coefficients $c_m(f)$ are given by (82.10).

Conversely, if $f : \mathbb{R} \to \mathbb{C}$ is a given 2π-periodic (Lipschitz continuous) function and we define $c_m(f)$ by (82.10), then we may ask if $f(x)$ can be represented by its Fourier series (82.11) for all x. We shall prove below that this is true if $f(x)$ is 2π-periodic and differentiable. This is the basic result of Fourier analysis stating that an arbitrary 2π-periodic differentiable function can be given a spectral decomposition in the form of a Fourier series. This result includes the "completeness" aspect of the basis functions $\{e^{imx}\}$, that is, the fact that *any* differentiable function can be represented as a Fourier series.

82.5 Fourier series as an orthonormal basis expansion

Normalizing the basis functions e^{-imx} we obtain the orthonormal basis basis functions $e_m(x) = \frac{1}{\sqrt{2\pi}}e^{-imx}$ satisfying

$$(e_m, e_k) = 0 \quad \text{if } k \neq m, \quad (e_m, e_m) = 1. \tag{82.12}$$

A Fourier series representation (**??**) takes the following form in the normalized basis:

$$f(x) = \sum_{m=-\infty}^{\infty} \tilde{c}_m(f)\frac{1}{\sqrt{2\pi}}e^{-imx}, \quad \tilde{c}_m(f) = \frac{1}{\sqrt{2\pi}}\int_{-\pi}^{\pi} f(x)e^{-imx}\,dx.$$

Of course, it would be natural to work with the normalized basis functions $\{\frac{1}{\sqrt{2\pi}}e^{-imx}\}$ and the corresponding renormalized Fourier coefficients $\tilde{c}_m(f)$ thus distributing the 2π-factor into two $\sqrt{2\pi}$-factors, but we follow the most common notation and include the 2π-factor in the Fourier coefficient $c_m(f)$ coupled to the basis function e^{imx}

82.6 Truncated Fourier series and best L_2-approximation

The *truncated Fourier series*

$$S_n f(x) = \sum_{m=-n}^{n} c_m(f)e^{imx}$$

of a given function $f(x)$ is a *best approximation* in the sense that

$$\|f - S_n f\| \leq \|f - g_n\|$$

for any function $g_n(x) = \sum_{m=-n}^{n} d_m e^{imx}$ with $d_m \in \mathbb{C}$, $m = 0, \pm 1, ..., \pm n$. This is because by the definition of the Fourier coefficents

$$(f - S_n f, e_m) = 0 \quad \text{for } m = 0, \pm 1, ..., \pm n,$$

and thus $S_n f(x)$ is the best approximation in the $L_2(-\pi, \pi)$ norm of $f(x)$ in the linear space spanned by the functions $\{1, e^{\pm ix}, e^{\pm i2x}, ..., e^{\pm inx}\}$, compare Chapter *Piecewise linear approximation*.

82.7 Real Fourier series: Cosine and Sine series representations

Using that $e^{imx} = \cos(mx) + i\sin(mx)$ and $\cos(-mx) = \cos(mx)$ and $\sin(-mx) = -\sin(mx)$, we can write (82.9) in the form

$$\sum_{m=-\infty}^{\infty} c_m e^{imx} = c_0 + \sum_{m=1}^{\infty} a_m \cos(mx) + \sum_{m=1}^{\infty} b_m \sin(mx),$$

where

$$a_m = c_m + c_{-m}, \quad b_m = i(c_m - c_{-m}), \quad m = 0, 1, 2, ...$$

If $f(x)$ is real, that is $f : \mathbb{R} \to \mathbb{R}$, then $\bar{c}_m = c_{-m}$ and thus $a_m = c_m + \bar{c}_m = 2\mathrm{Re}(c_m) \in \mathbb{R}$ and $b_m = i(c_m - \bar{c}_m) = 2\mathrm{Im}(c_n) \in \mathbb{R}$, and

$$c_m = \frac{a_m}{2} - i\frac{b_m}{2}, \quad c_{-m} = \frac{a_m}{2} + i\frac{b_m}{2} \quad , m = 0, 1, 2, ... \tag{82.13}$$

The Fourier series of a real-valued function 2π-periodic function $f : \mathbb{R} \to \mathbb{R}$ can thus be written alternatively as a Sine and Cosine series of the form

$$f(x) = \frac{a_0}{2} + \sum_{m=1}^{\infty} a_m \cos(mx) + \sum_{m=1}^{\infty} b_m \sin(mx),$$

where a_m, $b_m \in \mathbb{R}$, are given by

$$\begin{aligned} a_m = a_m(f) &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(mx)\,dx \quad \text{for } m = 0, 1, 2, ..., \\ b_m = b_m(f) &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(mx)\,dx \quad \text{for } m = 1, 2, ... \end{aligned}$$

We note that if $f(x)$ is even, that is $f(x) = f(-x)$, then $b_m = 0$ for $m = 1, 2, ...$, and thus $f(x)$ has a Cosine series representation

$$f(x) = \frac{a_0(f)}{2} + \sum_{m=1}^{\infty} a_m(f)\cos(mx). \tag{82.14}$$

Correspondingly, if $f(x)$ is odd, that is $f(x) = -f(-x)$, then $a_m = 0$ for $m = 0, 1, ...,$ and thus $f(x)$ a Sine series representation:

$$f(x) = \sum_{m=1}^{\infty} b_m(f) \sin(mx). \tag{82.15}$$

In the applications below we usually consider Cosine and Sine series for real-valued functions $f : \mathbb{R} \to \mathbb{R}$. The complex Fourier series is useful in the analysis of convergence of Fourier series.

We now present a couple of examples with Fourier coeffients having different rates of convergence to zero (m^{-2}, m^{-3} and m^{-1}).

EXAMPLE 82.1. Let $f : \mathbb{R} \to \mathbb{R}$ be a 2π-periodic function given by $f(x) = |x|$ for $-\pi \leq x \leq \pi$. The function $f(x)$ is real-valued and even, and thus has a Cosine series of the form (82.14). We compute using integration by parts if $m > 0$:

$$\begin{aligned}
a_0(f) &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\,dx = \frac{2}{\pi}\int_0^{\pi} x\,dx = \pi,\\
a_m(f) &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(mx)\,dx = \frac{2}{\pi}\int_0^{\pi} x\cos(mx)\,dx\\
&= \frac{2}{\pi}[\frac{x\sin(mx)}{m}]_0^{\pi} - \frac{2}{\pi}\int_0^{\pi}\frac{\sin(mx)}{m}\,dx = \frac{2}{\pi}\frac{(-1)^m - 1}{m^2}.
\end{aligned}$$

Since $(-1)^m - 1 = -2$ if m is odd and $(-1)^m - 1 = 0$ if m is even, the Fourier series representation of $f(x) = |x|$ takes the form

$$|x| = \frac{\pi}{2} - \frac{4}{\pi}\sum_{k=1}^{\infty}\frac{\cos((2k-1)x)}{(2k-1)^2}.$$

We plot the corresponding truncated series with summation over $k = 1, ..., n$ for different values of n in Fig. **??**:

A figure with truncated Cosine series of $f(x) = |x|$

EXAMPLE 82.2. Let $f : \mathbb{R} \to \mathbb{R}$ be an odd 2π-periodic function given by $f(x) = x(\pi - x)$ for $0 \leq x \leq \pi$. We compute its Sine series coefficients:

$$\begin{aligned}
b_m(f) &= \frac{1}{\pi}\int_{pi}^{\pi} f(x)\sin(mx)\,dx = \frac{2}{\pi}\int_0^{\pi} x(\pi - x)\sin(mx)\,dx\\
&= -\frac{2}{\pi}[\frac{x(\pi - x)\cos(mx)}{m}]_0^{\pi} + \frac{2}{\pi}\int_0^{\pi}\frac{(\pi - 2x)\cos(mx)}{m}\,dx\\
&= \frac{2}{\pi m}[\frac{(\pi - 2x)\sin(mx)}{m}]_0^{\pi} + \frac{2}{\pi m^2}\int_0^{\pi} 2\sin(mx)\,dx = \frac{4}{\pi m^3}(1 - (-1)^m).
\end{aligned}$$

Example 82.3. Define a 2π-periodic function $f(x)$ by setting

$$f(x) = \begin{cases} 1 & \text{for } |x| < a, \\ 0 & \text{for } a < |x| \leq \pi, \end{cases}$$

where $0 < a < \pi$, see Fig. **??**.

A fig of a piecewise constant 2π-periodic function and its truncated Fourier series

This is a piecewise Lipschitz continuous 2π-periodic even function, and we can compute its Fourier coefficients. We have $b_m(f) = 0$ and

$$2c_m(f) = a_m(f) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(mx)\,dx = \frac{2}{\pi}\int_0^a \cos(mx)\,dx = \frac{\sin(ma)}{m}.$$

We thus expect that $f(x) = \sum_{m=1}^{\infty} \frac{\sin(ma)}{m}\cos(mx)$. We shall return to this equality below, with particular focus on the values $x = \pm a$ where $f(x)$ has jump discontinuities.

82.8 Basic properties of Fourier coefficients

We now present some basic properties of the Fourier coefficients

$$c_m(f) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-imx}\,dx \quad m = 0, \pm 1, \pm 2, ...,$$

of a given 2π-periodic Lipschitz function $f : \mathbb{R} \to \mathbb{C}$.

82.8.1 Linearity

Fourier coefficients satisfy the following obvious linearity properties:

$$c_m(f+g) = c_m(f) + c_m(g), \quad c_m(\alpha f) = \alpha c_m(f),$$

where f and g are two functions with Fourier coeffcients $c_m(f)$ and $c_m(g)$, and $\alpha \in \mathbb{C}$.

82.8.2 Fourier coefficients of the derivative $Df = f'$

We now couple the Fourier coefficients of the derivative $Df = \frac{df}{dx}$ of a function 2π-periodic function $f : \mathbb{R} \to \mathbb{C}$ to the Fourier coefficients of f. The trick is to integrate by parts: Using the periodicity of $f(x)$, we find that

$$c_m(Df) = \frac{1}{2\pi}\int_{-\pi}^{\pi} Df(x)e^{-imx}\,dx = im\frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-imx}\,dx = im\,c_m(Df),$$

and we have thus proved:

Theorem 82.1 *If $f : \mathbb{R} \to \mathbb{C}$ is 2π-periodic and differentiable with derivative Df, then for $m = 0, \pm 1, \pm 2, ...$*

$$c_m(Df) = im\, c_m(f). \tag{82.16}$$

This is one of the fundamental results of Fourier analysis, and translates the operation of differentiation $D = \frac{d}{dx}$ with respect to x to multiplication of Fourier coefficients with im where m is the frequency. This opens the way of translating differential equations in the variable x to algebraic equations in the frequency m, which may be very useful and illuminating in certain applications.

We can directly generalize to

Theorem 82.2 *If $f : \mathbb{R} \to \mathbb{C}$ is 2π-periodic and k times differentiable with derivative $D^k f$, then for $m = 0, \pm 1, \pm 2, ...$*

$$c_m(D^k f) = (im)^k c_m(f). \tag{82.17}$$

EXAMPLE 82.4. Consider the differential equation $Du(x) + u(x) = f(x)$, where $f(x)$ a given 2π-periodic function and we seek a 2π-periodic solution $u(x)$. This equation models, for example, a resistor and capacitor in series, with $u(x)$ a primitive function of the current, $f(x)$ an applied voltage, and x repesenting time, see the Chapter Electrical circuits. Alernatively, $Du(x)+u(x) = f(x)$ models ain inductor and resistance in series with $u(x)$ now the current, and again $f(x)$ and applied voltage. For the Fourier coefficients we have using Theorem 82.1

$$im\, c_m(u) + c_m(u) = c_m(f),$$

and thus

$$c_m(u) = \frac{c_m(f)}{1 + im} = \frac{(1 - im)c_m(f)}{1 + m^2}.$$

This shows that the the indicated circuits act as so-called *low-pass filters*, with the property of damping high-frequency components: we view $f(x)$ as the input and $u(x)$ as the output and note that the Fourier coefficients of $u(x)$ decay quicker than those of $f(x)$.

EXAMPLE 82.5. Consider the differential equation $-D^2u(x) + u(x) = f(x)$ with $f(x)$ a given 2π-periodic function and we seek a 2π-periodic solution $u(x)$. Since $c_m D^2 u(m) = (im)^2 c_m(u)$, we obtain the following algebraic equation for the Fourier coefficients:

$$(m^2 + 1)c_m(u) = c_m(f) \quad \text{for } m \neq 0.$$

We can thus express the solution $u(x)$ of $-D^2u(x) = u(x) = f(x)$ as a Fourier series

$$u(x) = \sum_{-\infty}^{\infty} \frac{c_m(f)}{m^2 + 1} e^{imx},$$

if the data $f(x)$ is given as a Fourier series: $f(x) = \sum_{-\infty}^{\infty} c_m(f) e^{imx}$. Again, we see that the differential equation acts as a low-pass filter with damping of high-frequency components of the data $f(x)$.

EXAMPLE 82.6. More generally, consider the following differential equation $p(D)u(x) = f(x)$, where $p(D) = \sum_{k=0}^{p} a_k D^k$ is a differential equation with constant coefficients $a_k \in \mathbb{C}$, the data $f(x)$ is 2π-periodic and we seek a 2π-periodic solution $u(x)$. Arguing as above, we get the following equation for the Fourier coefficients:

$$p(im)c_m(u) = \sum_{k=0}^{p} a_k (im)^k c_m(u) = c_m(f),$$

that is, assuming $p(im) \neq 0$ (or $c_m(f) = 0$ if $p(im) = 0$),

$$c_m(u) = \frac{c_m(f)}{p(im)},$$

which gives the Fourier series for the solution, if the the Fourier series for the data $f(x)$ is given.

82.9 The Fourier coefficients $c_m(f)$ tend to zero as $|m| \to \infty$

As a direct consequence of the preceeding result, we conclude that the Fourier coefficients $c_m(f)$ of a 2π-periodic differentiable function $f(x)$ with integrable derivative Df, tend to zero as $|m|$ tends to infinity: Since $|im\, c_m(f)| = |c_m(Df)|$, we have

$$|c_m(f)| = \frac{1}{|m|}|c_m(Df)| \leq \frac{1}{|m|} \int_{-\pi}^{\pi} |Df|\, dx \to 0 \quad \text{as } |m| \to \infty.$$

Similarly, if $f(x)$ is 2π-periodic with integrable derivative $D^k f$ of order $k > 1$, then for $m = \pm 1, \pm 2, ...,$

$$|c_m(f)| \leq \frac{1}{|m^k|} \int_{-\pi}^{\pi} |D^k f|\, dx.$$

We conclude that the larger k is, the more rapid is the convergence of $c_m(f)$ to zero.

We can also go in the direction of less regularity and ask if we can show that the Fourier coefficients $c_m(f)$ tend to zero as $|m| \to \infty$ under the weaker assumption that f is Lipschitz continuous only. To this end we first note that for and $-\pi < a < b < \pi$, we have

$$\int_a^b e^{-imx} dx = \frac{1}{-im}[e^{-imx}]_a^b \to 0 \quad \text{as } m \to 0. \tag{82.18}$$

This may be seen as a consequence of the rapid oscillations of e^{imx} with $|m|$ large, which causes a lot of cancellations in any integral of the form (82.18) with the effect that the integrals decreases to zero as m increases to infinity, see the following fig

An illustration of the fact that $\int_a^b \cos(mx)dx$ and $\int_a^b \sin(mx)dx$ is small for m large

The estimate (82.18) shows that if $f(x)$ is piecewise constant on $[-\pi, \pi]$, that is a linear combination sum of functions equal to one on a certain interval and zero elsewhere, then $c_m(f) \to 0$ as $|m| \to \infty$.

Finally, a given Lipschitz continuous function $f : [-\pi, \pi] \to \mathbb{C}$ can be approximated by a piecewise constant function $\tilde{f}(x)$, so that

$$\int_{-\pi}^{\pi} |f(x) - \tilde{f}(x)|$$

is a s small as we please, which leads to the famous

Theorem 82.3 (Riemann-Lebesgues lemma) *If $f : [-\pi, \pi]$ is Lipschitz continuous, then $c_m(f) \to 0$ as $|m| \to \infty$.*

The assumption can be relaxed to piecewise Lipschitz continuity.

82.10 Convolution

Given two 2π-periodic functions $f(x)$ and $g(x)$, we define a new 2π-periodc function $f * g$ by

$$(f * g)(x) = \int_{-\pi}^{\pi} f(x-y)g(y)\, dy \quad x \in \mathbb{R}.$$

We say that $f * g$ is the *convolution* of f and g. Changing variables, setting $y = x - t$, we find that

$$(f * g)(x) = \int_{-\pi}^{\pi} f(t)g(x-t)\, dt = \int_{-\pi}^{\pi} f(y)g(x-y)\, dy \quad x \in \mathbb{R},$$

and thus the integrand can take the form $f(x-y)g(y)$ or $f(y)g(x-y)$.

We shall now prove that

$$c_m(f * g) = 2\pi\, c_m(f)c_m(g). \tag{82.19}$$

By direct computation we have changing order of integration and using the change of variable $t = x - y$

$$
\begin{aligned}
c_m(f*g) &= \frac{1}{2\pi}\int_{-\pi}^{\pi}(f*g)(x)e^{-imx}\,dx = \frac{1}{2\pi}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(x-y)g(y)\,dy e^{-imx}\,dx \\
&= \int_{-\pi}^{\pi} g(y)e^{-imy}(\frac{1}{2\pi}\int_{-\pi}^{\pi} f(x-y)\,dy e^{-im(x-y)}\,dx)dy \\
&= \int_{-\pi}^{\pi} g(y)e^{-imy}(\frac{1}{2\pi}\int_{-\pi}^{\pi} f(t)\,e^{-imt}\,dt)dy \\
&= c_m(f)\int_{-\pi}^{\pi} g(y)e^{-imy}\,dy = 2\pi c_m(f)c_m(g).
\end{aligned}
$$

EXAMPLE 82.7. Let $g : \mathbb{R} \to$ be a 2π-periodic function defined by

$$
g(x) = \frac{1}{2a} \quad \text{for } -a \leq x \leq a
$$

where $0 < a < \pi$. For a small, we may view $g(x)$ as an approximate delta function. The convolution

$$
f * g(x) = \int_{-\pi}^{\pi} f(x-y)g(y)\,dy = \int_{-a}^{a} f(x-y)\,dy
$$

is an average of $f(x)$ over the interval $[x-a, x+a]$. Recalling (**??**), and using (82.19), we get

$$
c_m(f*g) = c_m(f)\frac{\sin(ma)}{ma}.
$$

We conclude that $c_m(f*g)$ is close to $c_m(f)$ if ma is small, and $c_m(f*g)$ is much smaller than $c_m(f)$ if ma is large. The Fourier coefficients of the average $f*g$ thus decay quicker than those of f, and thus $f*g$ is a smoothed version of f: taking the average increases smoothness reflected by quickly decreasing Fourier coefficients.

82.11 The inversion formula

We shall now prove that if $f : \mathbb{R} \to \mathbb{C}$ is 2π-periodic and differentiable, then for all $x \in \mathbb{R}$

$$
\lim_{n\to\infty}\sum_{-n}^{n} c_m(f)e^{imx} = f(x).
$$

In other words, the function $f(x)$ can be represented as a convergent Fourier series:

$$
f(x) = \sum_{m=-\infty}^{\infty} c_m(f)e^{imx} \quad \text{for } x \in \mathbb{R}.
$$

We have

$$\sum_{-n}^{n} c_m e^{imx} = \sum_{-n}^{n} \frac{1}{2\pi} \int_{-\pi}^{\pi} f(y) e^{-imy}\, dy\, e^{imx}$$
$$= \int_{-\pi}^{\pi} f(y) \frac{1}{2\pi} \sum_{-n}^{n} e^{im(x-y)}\, dy = \int_{-\pi}^{\pi} f(y) D_n(x-y)\, dy, \tag{82.20}$$

where, setting $\theta = x - y$,

$$D_n(\theta) = \frac{1}{2\pi} \sum_{-n}^{n} e^{im\theta} = \frac{1}{2\pi} e^{-in\theta} \sum_{m=0}^{2n} e^{im\theta}$$
$$= \frac{1}{2\pi} e^{-in\theta} \frac{1 - e^{i(2n+1)\theta}}{1 - e^{i\theta}} = \frac{1}{2\pi} \frac{e^{-i\frac{\theta}{2}}}{e^{-i\frac{\theta}{2}}} \frac{e^{-in\theta} - e^{i(n+1)\theta}}{1 - e^{i\theta}} = \frac{1}{2\pi} \frac{\sin(n\theta + \frac{\theta}{2})}{\sin(\frac{\theta}{2})}$$

is the so-called *Dirichlet kernel.* We here used that $\sum_{m=0}^{2n} e^{im\theta}$ is a finite geometric series with factor $e^{i\theta}$. Using the convolution notation we can write (83.4) in the compact form

$$\sum_{-n}^{n} c_m e^{imx} = f * D_n(x).$$

In order for $f * D_n(x)$ to approximate $f(x)$, we expect D_n to somehow behave like the identity. We look a plot of $D_n(\theta)$:

A plot of $D_n(\theta)$

We see that $D_n(\theta)$ oscillates and has a peak at $\theta = 0$. Integrating $D_n(\theta) = \frac{1}{2\pi} \sum_{-n}^{n} e^{im\theta}$ term by term over $[-\pi, \pi]$ noting that all integrated terms vanish but one, we see that the total area (with sign) under the graph of D_n is equal to one, that is

$$\int_{-\pi}^{\pi} D_n(\theta)\, d\theta = 1, \tag{82.21}$$

which expresses one aspect of the idea that D_n behaves like the identity. The other aspect of the approximate identity nature of D_n is the increasing focussing of the peak of D_n at 0 as n increases.

Using (83.5), we can write

$$f(x) - f * D_n(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} (f(x) - f(y)) D_n(x-y)\, dy$$
$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} g(x,y) \sin((n + \frac{1}{2})(x-y))\, dy$$

where

$$g(x, y) = \frac{f(x) - f(y)}{\sin(\frac{x-y}{2})}.$$

Now if $f(x)$ is twice differentiable, then $g(x, y)$ is differentiable with respect to y for all $y \in \mathbb{R}$ with derivative $Dg(x, y)$ (see the corresponding argument in the proof of Cauchy's formula). Integrating by parts we thus have

$$f(x) - D_n * f(x) = -\frac{1}{2\pi}\frac{1}{n+\frac{1}{2}}\int_{-\pi}^{\pi} Dg(x, y)\cos((n+\frac{1}{2})(x-y))\,dy \to 0$$

as $n \to \infty$. In case $f(x)$ is differentiable with piecewise Lipschitz continuous derivative, then $Dg(x, y)$ is Lipschitz continuous in y, and by Riemann-Lebesgues' lemma, we find the same conclusion. We summarize in the following basic theorem:

Theorem 82.4 *If $f : \mathbb{R} \to \mathbb{C}$ is 2π-periodic with piecewise Lipschitz continuous derivative, then $f(x)$ may be represented by a convergent Fourier series:*

$$f(x) = \sum_{m=-\infty}^{\infty} c_m(f)e^{imx} \quad \textit{for } x \in \mathbb{R},$$

where the coefficients c_m are given by (82.10).

The assumption on $f(x)$ can be relaxed: it suffices to assume that $f(x)$ is piecewise differentiable with piecewise Lipschitz continuous derivative. At a point x of discontinuity, the Fourier series converges to the mean value of the left hand limit $f^-(x) = \lim_{y\to x, y<x} f(y)$and the right hand limit $f^+(x) = \lim_{y\to x, y>x} f(y)$:

$$\sum_{m=-\infty}^{\infty} c_m(f)e^{imx} = \frac{f^-(x) + f^+(x)}{2}. \tag{82.22}$$

EXAMPLE 82.8. We have

$$\sum_{m=1}^{\infty}\frac{sin(ma)}{ma}\cos(mx) = \begin{cases} 1 & \text{if } |x| < a, \\ \frac{1}{2} & \text{if } |x| = a, \\ 0 & \text{if } |x| > a. \end{cases}$$

82.12 Parsevals formula

Suppose $f : \mathbb{R} \to \mathbb{R}$ is 2π periodic with a convergent Fourier series representation:

$$f(x) = \sum_{m=-\infty}^{\infty} c_m(f)e^{imx}, \tag{82.23}$$

where

$$c_m(f) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-imx}\,dx.$$

Using the orthogonality (82.7) of the functions $\{e^{imx}\}$, we find that

$$\int_{-\pi}^{\pi} |f(x)|^2\,dx = \int_{-\pi}^{\pi} f(x)\overline{f(x)}\,dx = \int_{-\pi}^{\pi} (\sum_{m=-\infty}^{\infty} c_m(f)e^{imx})(\sum_{k=-\infty}^{\infty} \overline{c_k(f)}e^{-ikx})\,dx$$

$$\sum_{m,k=-\infty}^{\infty} c_m(f)\overline{c_k(f)}\int_{-\pi}^{\pi} e^{imx}e^{-ikx}\,dx = 2\pi\sum_{m=-\infty}^{\infty} |c_m(f)|^2.$$

We have now proved the celebrated:

Theorem 82.5 (Parseval's formula) *If $f(x)$ has a convergent Fourier series representation, then*

$$\int_{-\pi}^{\pi} |f(x)|^2\,dx = 2\pi\sum_{m=-\infty}^{\infty} |c_m(f)|^2.$$

We can in an obvious way generalize to

Theorem 82.6 *If $f(x)$ and $g(x)$ have convergent Fourier series representations, then*

$$\int_{-\pi}^{\pi} f(x)\overline{g(x)}\,dx = 2\pi\sum_{m=-\infty}^{\infty} c_m(f)\overline{c_m(g)}.$$

82.13 Space vs frequency analysis

We are now ready to lean back an reflect a bit about the nature of Fourier series. Suppose that $f(x)$ is a given 2π-periodic function. If we want to describe the nature of the function $f(x)$, that is the variation of $f(x)$ with x, we can try to give some kind of list of $f(x)$ different for values x. We may call this physical description where we think of x as a space or time variable. Now using Fourier series we can instead express the function $f(x)$ as a Fourier series, determined by the Fourier coefficients $\{c_m(f)\}$. Describing $f(x)$ through its Fourier coefficients, may be viewed as a frequency-description. In the physical description, we describe the function f in terms of its function values $f(x)$ for different values of x. In the frequency description, we describe f in terms of the Fourier coefficients $c_m(f)$ as a sum $f(x) = \sum_m c_m(f)e^{imx}$.

To describe a given function $f(x)$ we may thus look at the variation of $f(x)$ with x, or the the variation of $c_m)f)$ with m.

We have noted that the decay of $c_m(f)$ with m couples to the regularity of $f(x)$:if $f(x)$ is highly regular with many derivatives, then the Fourier coefficients $c_m(f)$ decay quickly with increasing m, and vice versa. If the Fourier coefficients decay quickly, then only a few terms in the Fourier series suffices to represent the function to high accuracy.

82.14 Fourier series of periodic functions with period different from 2π

Suppose $f : \mathbb{R} \to \mathbb{C}$ is periodic with period $\frac{2\pi}{\omega}$ with $\omega > 0$. We considered above the case $\omega = 1$, and we now generalize to $\omega > 0$. For example: the functions $\sin(\omega x)$, $\sin(2\omega x)$, $\sin(3\omega x)$..., are periodic with period $\frac{2\pi}{\omega}$.

Defining $g(x) = f(\frac{x}{\omega})$, we have that $g(x)$ is 2π-periodic since $g(x+2\pi) = f(\frac{x+2\pi}{\omega}) = f(\frac{x}{\omega} + \frac{2\pi}{\omega}) = f(\frac{x}{\omega}) = g(x)$, and a Fourier series representation of $g(x)$:

$$g(x) = \sum_{m=-\infty}^{\infty} c_m(g) e^{imx}, \quad c_m(g) = \frac{1}{2\pi} \int_{-\pi}^{\pi} g(y) e^{-imy}\, dy$$

translates into the following Fourier series representation of $f(\frac{x}{\omega})$:

$$f(\frac{x}{\omega}) = \sum_{m=-\infty}^{\infty} c_m(g) e^{imx}, \quad c_m(g) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\frac{y}{\omega}) e^{-imy}\, dy$$

which takes the following form changing variables from $\frac{x}{\omega}$ to x and $\frac{y}{\omega}$ to y:

$$f(x) = \sum_{m=-\infty}^{\infty} c_m(f) e^{im\omega x}, \quad c_m(f) = \frac{\omega}{2\pi} \int_{-\frac{\pi}{\omega}}^{\frac{\pi}{\omega}} f(y) e^{-im\omega y}\, dy. \qquad (82.24)$$

EXAMPLE 82.9.

82.15 Weierstrass functions

Consider a series of the form

$$\sum_{m=1}^{\infty} a^{-m} \sin(b^m x), \qquad (82.25)$$

where $a > 1$, $b > a$. This type of series was presented by Weierstrass as an example of a Lipschitz continuous function that is not differentiable at any point, see Fig. 82.25, where we also plot the corresponding truncated series $\sum_{m=1}^{n} a^{-m} \sin(b^m x)$ for different values of n. We see that as n increases the series oscillates increasingly wildly, and gives an irregular "chaotic" impression.

Since $a > 1$ the series (82.25) is absolutely convergent, and defines a function $f(x) = \sum_{m=1}^{\infty} a^{-m} \sin(b^m x)$, but the series

$$\sum_{m=1}^{\infty} a^{-m} b^m \cos(b^m x)$$

obtained by termwise differentition, does not converge since $\frac{b}{a} > 1$, which indicates that $f(x)$ is nowhere differentiable. The Weierstrass function, or the corresponding truncated series, is an example of a function with a seuqence of "microscales" $\frac{2\pi}{b^m}$, $m = 1, 2, ...$ corresponding to the different basis functions $\sin(b^m x)$. The function $f(x)$ thus has the same oscillating nature on all scales and thus has a "fractal" nature. It is believed that phenomena like turbulence also have a fractal nature, which may be useful in attempts to model microscales which are not possible to model numerically.

Plots of Weierstrass functions

Choosing $b = 2$ (or b any natural number > 1), gives a series of the form $\sum_{m=1}^{\infty} a^{-m} \sin(2^m x)$, which is an example of a *lacunary Fourier series*, with just very few Fourier coefficients being non-zero. A Weierstrass function with b a natural number is thus a lacunary Fourier series.

82.16 Solving the 1d heat equation using Fourier series

We consider the 1d homogeneuous heat equation:

$$\begin{aligned} \dot{u}(x,t) - u''(x,t) &= 0 \quad \text{for } 0 < x < \pi,\ t > 0, \\ u(0,t) = u(\pi,t) &= 0 \quad \text{for } t > 0, \\ u(x,0) &= u_0(x) \quad \text{for } 0 < x < \pi, \end{aligned} \tag{82.26}$$

where u_0 is a given initial value. We observe that for $m = 1, 2, ...,$, the function $v(x,t) = e^{-m^2 t} \sin(mx)$ satisfies

$$\dot{v}(x,t) - v''(x,t) = 0 \quad \text{for } 0 < x < \pi, \quad v(0,t) = v(\pi,t) = 0 \quad \text{for } t > 0,$$

and thus any finite linear combination

$$u(x,t) = \sum_{m=1}^{J} b_m e^{-m^2 t} \sin(mx)$$

with coefficinets $b_m \in \mathbb{R}$, satisfies (83.6) with corresponding initial data $u_0 = \sum_{m=1}^{J} b_m \sin(mx)$. Each term $e^{-m^2 t} \sin(mx)$ has the form of a product of a function of x only , namely $\sin(mx)$ with *frequency* m, and a

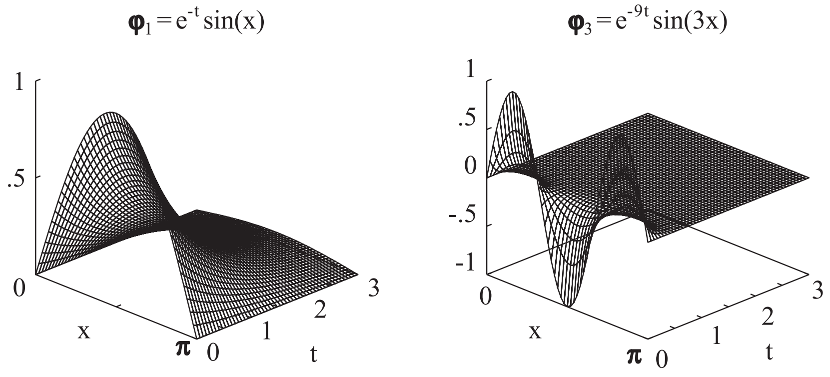


FIGURE 82.2. The solutions of the heat equation corresponding to frequencies $j = 1$ and $j = 3$.

function of t only, namely $e^{-m^2 t}$. We see that the factor $e^{-m^2 t}$ decays with increasing t and the rate of decay increases quickly with increasing frequency m. We illustrate this in Fig. 82.2.

More generally, if the initial data u_0 has a convergent Sine series (with $u_0(x)$ extended as an odd function to $[-\pi, \pi]$)

$$u_0(x) = \sum_{m=1}^{\infty} b_m(u_0) \sin(mx), \tag{82.27}$$

with Fourier coefficients

$$b_m(u_0) = \frac{2}{\pi} \int_0^{\pi} u_0(x) \sin(mx)\, dx, \tag{82.28}$$

then the function defined by

$$u(x,t) = \sum_{m=1}^{\infty} b_m(u_0) e^{-m^2 t} \sin(mx), \tag{82.29}$$

solves the intitial value problem (83.6).

82.17 Computing Fourier coefficients with quadrature

To compute the Fourier coefficients

$$c_m(f) = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-imx}\, dx, m = 0, \pm 1, \pm 2, ...$$

of a given 2π-periodic function $f:\mathbb{R}\to\mathbb{C}$, we will in general have to use quadrature. Using the quadrature points $x_n=\frac{2\pi n}{N}$, $n=0,...,N-1$, with weights $\omega_n=\frac{2\pi}{N}$, corresponding to a left end-point quadrature formula with N uniformly distributed points, we would approximate $c_m(f)$ by

$$c_m(f)=\frac{1}{2\pi}\int_0^{2\pi} f(x)e^{-imx}\,dx\approx\frac{1}{2\pi}\sum_{n=0}^{N-1} f(x_n)e^{-imx_n}\,\omega_n\equiv\widehat{f}(m),$$

We cannot expect this quadrature formula to be accurate for $m>N$ since then the variation of e^{imx} would not be captured by the quadrature points $\frac{2\pi n}{N}$. We note that $\widehat{f}(m)$ is periodic with period N: that is $\widehat{f}(m)=\widehat{f}(m+N)$, and it it thus natural to consider $\widehat{f}(m)$ for $m=0,...,N-1$, or equivalently with $|m|\leq(N-1)/2$. We call $(N-1)/2$ the *Nyquist cut-off frequency* which corresponds to at least 2 quadrature points on each period for frequencies m with $|m|\leq(N-1)/2$. According to the inversion formula, we could hope, assuming that the Fourier coefficients $c_m(f)$ are small enough for m larger than cut-off, that

$$f(x_n)\approx\sum_{m=0}^{N-1}\widehat{f}(m)e^{imx_n}\quad\text{for } n=0,...,N-1,\tag{82.30}$$

which thus would represent an approximate discrete Fourier decomposition for the selected values x_n based on computing the Fourier coefficients $c_m(f)$ by quadrature for $m=0,...,N-1$. This leads us directly into the *discrete Fourier transform*, which we now discuss.

82.18 The Discrete Fourier Transforms

Suppose $\{f_n\}_{n=0}^{N-1}$ is a finite sequence of N given complex numbers. We define a corresponding sequence $\{\widehat{f}_m\}_{m=0}^{N-1}$ by

$$\widehat{f}_m=\frac{1}{N}\sum_{n=0}^{N-1} f_n e^{-2\pi imn/N},\quad\text{for } m=0,...,N-1.$$

We say that the sequence $\{\widehat{f}_m\}_{m=0}^{N-1}$ is the *discrete Fourier transform* of the sequence $\{f_n\}_{n=0}^{N-1}$. In the setting of the previous section we have $f_n=f(\frac{2\pi n}{N})$ and $\widehat{f}_m\approx c_m(f)$.

We find from the definitions

$$\begin{aligned}&=\sum_{m=0}^{n-1}\widehat{f}(m)e^{2\pi imn/N}=\sum_{m=0}^{n-1}\frac{1}{N}\sum_{k=0}^{N-1} f_k e^{-2\pi imk/N}e^{2\pi imn/N}\\&=\sum_{k=0}^{N-1} f_k\frac{1}{N}\sum_{m=0}^{n-1} e^{2\pi im(n-k)/N}\end{aligned}$$

and using that

$$\frac{1}{N}\sum_{m=0}^{n-1} e^{2\pi i m(n-k)/N} = \begin{cases} 1 & \text{if } k = n, \\ 0 & \text{else,} \end{cases}$$

we obtain the following inversion formula, to be compared with (82.30),

$$f_n = \sum_{m=0}^{N-1} \hat{f}(m) e^{2\pi i m n/N}, \quad \text{for } n = 0, ..., N-1. \tag{82.31}$$

To compute the discrete Fourier transform of $\{f_n\}_{n=0}^{N-1}$, we would need on the order of N^2 operations. If $N = 2^k$ for som natural number k, it is possible to organize the computation of the discrete Fourier transform so that required operations would be of the order N up to a logarithm. The corresponding transform referred to as the *Fast Fourier Transform FFT* developed by Cooley and Tukey in the 1960s, is one of the highlights of applied mathematics of modern time.

Chapter 82 Problems

82.1. Complete the details of the proof of (82.17).

82.2. Prove (82.22).

82.3. Some technicalities: translation, scaling...

82.4. Show that the Sine series coefficients for the odd function $f(x) = x^3 - \pi^2 x$ for $-\pi \leq x \leq \pi$, are given by $b_m(f) = 12\frac{(-1)^m}{m^3}$.

82.5. Show that the Cosine series coefficients for the even function $f(x) = x^4 - 2\pi^2 x^2$ for $-\pi \leq x \leq \pi$, are given by $a_0 = \frac{14\pi^4}{15}$, $a_m(f) = 48\frac{(-1)^{m+1}}{m^4}$, $m = 1, 2, ...$

82.6. Prove that $\sum_{m=1}^{\infty} \frac{1}{m^4} = \frac{\pi^4}{90}$.

83 Fourier transforms

> As the natural ideas of equality developed it was possible to conceive the sublime hope of establishing among us a free government exempt from kings and priests, and to free from this double yoke the long-usurped soil of Europe. I readily became enamoured of this cause, in my opinion the greatest and most beautiful which any nation has ever undertaken. (Fourier 1793, joining a Revolutionary Committee of the French Revolution)

Fourier series concern function $f : \mathbb{R} \to \mathbb{C}$ which are periodic. We now consider functions $f : \mathbb{R} \to \mathbb{C}$ which are non-periodic and the analogous concept is then the *Fourier transform*, which we will study in this chapter. For a given (piecewise Lipschitz continuous) function $f : \mathbb{R} \to \mathbb{C}$ such that $f(x)$ is integrable over $\mathbb{R}$, that is,

$$\int_{\mathbb{R}} |f(x)|\, dx < \infty, \tag{83.1}$$

we define for $\xi \in \mathbb{R}$

$$\widehat{f}(\xi) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(x) e^{-i\xi x}\, dx, \tag{83.2}$$

noting that the integral is absolutely convergent and thus well defined under the assumption (83.1). We say that the function $\widehat{f} : \mathbb{R} \to \mathbb{C}$ defined by (83.2) is the *Fourier transform* of $f(x)$.

We shall now develop a calculus for the Fourier transform which is analogous to that developed for Fourier series in the previous chapter. In particular we shall prove the inversion formula:

$$f(x) = \int_{-\infty}^{\infty} \widehat{f}(\xi)e^{i\xi x}\, d\xi \quad \text{for } x \in \mathbb{R},$$

under the assumption that $f(x)$ is differentiable on $\mathbb{R}$. As we go along the analogy between Fourier series and Fourier transforms will be uncovered.

We compute the Fourier transform of some basic functions.

EXAMPLE 83.1. If $f(x) = e^{-|x|}$ for $x \in \mathbb{R}$, then

$$\begin{aligned}
\widehat{f}(\xi) &= \frac{1}{2\pi}\int_{\infty}^{\infty} e^{|x|-i\xi x}\, dx \\
&= \frac{1}{2\pi}\int_{-\infty}^{0} e^{-x-i\xi x}\, dx + \frac{1}{2\pi}\int_{0}^{\infty} e^{x-i\xi x}\, dx \\
&= \frac{1}{2\pi}\frac{1}{1+i\xi} + \frac{1}{2\pi}\frac{1}{1-i\xi} = \frac{1}{\pi}\frac{1}{1+\xi^2}.
\end{aligned}$$

EXAMPLE 83.2. If $f(x) = e^{-\frac{ax^2}{2}}$ for $x \in \mathbb{R}$, where $a > 0$ is a constant then

$$\widehat{f}(\xi) = \frac{1}{2\pi}\int_{\infty}^{\infty} e^{-\frac{ax^2}{2}} e^{-i\xi x}\, dx = \frac{1}{2\pi}e^{-\frac{\xi^2}{2a}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(\sqrt{a}x + i\frac{\xi}{\sqrt{a}})^2}\, dx.$$

To evaluate the integral, we shall use Cauchy's theorem for analytic functions as follows: We note that the function $g(z) = e^{-\frac{1}{2}(\sqrt{a}z+i\frac{\xi}{\sqrt{a}})^2}$ is analytic in z, and we may thus shift the line of integration to obtain

$$\widehat{f}(\xi) = \frac{1}{2\pi}e^{-\frac{\xi^2}{2a}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(\sqrt{a}x)^2}\, dx,$$

and recalling (64.25) we thus have

$$\widehat{f}(\xi) = \frac{1}{2\pi}e^{-\frac{\xi^2}{2a}}\int_{-\infty}^{\infty} e^{-\frac{ax^2}{2}}\, dx = \frac{1}{2\sqrt{a}}e^{-\frac{\xi^2}{2a}}$$

We note that as a tends to zero, the function $f(x)$ tends to 1 for all x, and $\widehat{f}(\xi)$ tends to $\delta(0)$ the delta function at 0, see Fig. **??**

fig of $e^{-\frac{ax^2}{2}}$ and its Fourier transform $\frac{1}{2\sqrt{a}}e^{-\frac{\xi^2}{2a}}$ for different $a > 0$

EXAMPLE 83.3. Defining $f(x)$ by

$$f(x) = 1 \quad \text{for } -a \leq x \leq a$$

where $a > 0$, and $f(x) = 0$ else, we obtain (see Fig. **??**

$$\widehat{f}(\xi) = \frac{1}{2\pi}\int_{-a}^{a} e^{-i\xi x}\, dx = \frac{\sin(\xi a)}{\xi}.$$

Fig of above function and its Fourier transform

83.1 Basic properties of the Fourier transform

We now present some basic properties of the Fourier transform

$$\widehat{f}(\xi) = \frac{1}{2\pi}\int_{\infty}^{\infty} f(x)e^{-i\xi x}\,dx, \quad \xi \in \mathbb{R},$$

of given functions $f : \mathbb{R} \to \mathbb{R}$ which are integrable over $\mathbb{R}$.

83.1.1 Linearity

The Fourier transform satisfies the following obvious linearity properties:

$$\widehat{(f+g)}(\xi) = \widehat{f}(\xi) + \widehat{g}(\xi), \quad \widehat{(\alpha f)}(\xi) = \alpha\widehat{f}(\xi).$$

where f and g are two functions with Fourier transforms $\widehat{f}$ and $\widehat{g}$, and $\alpha \in \mathbb{C}$.

83.1.2 Scaling

Let $g : \mathbb{R} \to \mathbb{R}$ be integrable and define $f(x) = g(ax)$, where $a > 0$ is a constant. Then, changing variables setting $y = ax$, we have

$$\hat{f}(\xi) = \frac{1}{2\pi}\int_{\infty}^{\infty} g(ax)e^{-i\frac{\xi}{a}ax}\,dx = \frac{1}{2\pi}\int_{\infty}^{\infty} g(y)e^{-i\frac{\xi}{a}y}\,\frac{1}{a}dy = \frac{1}{a}\hat{g}(\frac{\xi}{a}).$$

We conclude that if $f(x) = g(ax)$, then $\hat{f}(\xi) = \frac{1}{a}\hat{g}(\frac{\xi}{a})$.

83.1.3 The Fourier transform of the derivative $Df = \frac{df}{dx}$

We now couple the Fourier transform of the derivative $Df = \frac{df}{dx}$ of a function f to the Fourier transform of f. The trick is to integrate by parts:

$$\widehat{Df}(\xi) = \frac{1}{2\pi}\int_{-\infty}^{\infty} Df(x)e^{-i\xi x}\,dx = i\xi\frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)e^{-i\xi x}\,dx = i\xi\widehat{f}(\xi).$$

We summarize in the following theorem:

Theorem 83.1 *If $f : \mathbb{R} \to \mathbb{C}$ is integrable with integrable derivative Df, then for $\xi \in \mathbb{R}$,*

$$\widehat{Df}(\xi) = i\xi\widehat{f}(\xi) \tag{83.3}$$

This is one of the fundamental results of Fourier analysis, and translates the operation of differentiation $D = \frac{d}{dx}$ with respect to x to multiplication of Fourier coefficients with $i\xi$ where ξ is the frequency. This opens the way of translating differential equations in the variable x to algebraic equations in the frequency ξ, which may be very useful and illuminating in certain applications.

EXAMPLE 83.4. Consider the differential equation $-D^2u(x) + u(x) = f(x)$ on $\mathbb{R}$ with $f(x)$ a given integrable function and seeking an integrable solution $u(x)$. Since $\widehat{D^2u}(\xi) = (i\xi)^2\widehat{u}(\xi)$, we obtain the algebraic equation

$$(\xi^2 + 1)\widehat{u}(\xi) = \widehat{f}(\xi) \quad \text{for } \xi \neq 0$$

and we can thus express the solution $u(x)$ as a Fourier integral

$$u(x) = \int_{\mathbb{R}} \frac{\widehat{f}(\xi)}{\xi^2 + 1} e^{i\xi x}\, d\xi,$$

in terms of the Fourier transform $\widehat{f}(\xi)$ of the data $f(x)$.

83.2 The Fourier transform $\widehat{f}(\xi)$ tends to zero as $|\xi| \to \infty$

As a direct consequence of the preceeding result, we conclude that the Fourier transform $\widehat{f}(\xi)$ of a differentiable function $f(x)$ with integrable derivative $Df(x)$, tend to zero as $|\xi|$ tends to infinity. This is simply because

$$|\widehat{f}(\xi)| = \frac{1}{|\xi|}\widehat{f}(\xi) \leq \frac{1}{|\xi|}\int_{-\infty}^{\infty} |Df|\, dx \to 0 \quad \text{as } |\xi| \to \infty.$$

. This result can be extended to the case of $f(x)$ being integrable, as in the corresponding case of Fourier series.

83.2.1 Convolution

Given two functions $f : \mathbb{R} \to \mathbb{R}$ and $g\mathbb{R} \to \mathbb{R}$, we define a new function $f * g : \mathbb{R} \to \mathbb{R}$ by

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x - y)g(y)\, dy.$$

We say that $f * g$ is the *convolution* of f and g. We shall prove that

$$\widehat{f * g} = 2\pi\widehat{f}\widehat{g}$$

By direct computation we have changing order of integration and using the change of variable $t = x - y$

$$\begin{aligned}
\widehat{f * g}(\xi) &= \frac{1}{2\pi}\int_{-\infty}^{\infty}(f*g)(x)e^{-i\xi x}\,dx = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x-y)g(y)\,dy e^{-i\xi x}\,dx \\
&= \int_{-\infty}^{\infty} g(y)e^{-i\xi y}(\frac{1}{2\pi}\int_{-\infty}^{\infty} f(x-y)\,dy e^{-i\xi(x-y)}\,dx)dy \\
&= \int_{-\infty}^{\infty} g(y)e^{-i\xi y}(\frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)\,dy e^{-i\xi t}\,dt)dy \\
&= \widehat{f}(\xi)\int_{-\infty}^{\infty} g(y)e^{-i\xi y}\,dy = 2\pi\widehat{f}(\xi)\widehat{g}(\xi).
\end{aligned}$$

We summarize:

Theorem 83.2 *We have* $\widehat{f * g}(\xi) = 2\pi\widehat{f}(\xi)\widehat{g}(\xi)$ *for* $\xi \in \mathbb{R}$.

EXAMPLE 83.5.

83.3 The inversion formula

We shall now prove that if $f(x)$ is differentiable, then for all $x \in \mathbb{R}$,

$$\lim_{n\to\infty} f_n(x) = f(x),$$

where

$$f_n(x) = \int_{-n}^{n}\widehat{f}(\xi)e^{i\xi x}\,d\xi$$

and thus for all $x \in \mathbb{R}$, the function $f(x)$ can be represented as a convergent Fourier integral:

$$f(x) = \int_{-\infty}^{\infty}\widehat{f}(\xi)e^{i\xi x}\,d\xi.$$

We have

$$f_n(x) = \int_{-n}^{n}\widehat{f}(\xi)e^{i\xi x}\,d\xi = \frac{1}{2\pi}\int_{\infty}^{\infty} f(y)\int_{-n}^{n} e^{i\xi(x-y)}\,d\xi dy = \int_{\infty}^{\infty} f(y)D_n(x-y)\,dy, \tag{83.4}$$

where, setting $\theta = x - y$,

$$D_n(\theta) = \frac{1}{2\pi}\int_{-n}^{n} e^{i\xi\theta}\,d\xi = \frac{1}{\pi}\frac{\sin(n\theta)}{\theta}$$

is the *Dirichlet kernel* for the Fourier transform. Using the convolution notation we can write (83.4) in the compact form

$$f_n(x) = f * D_n(x).$$

With experience from the Dirichlet kernel for Fourier series, we expect D_n to be an approximate identity. Looking at a plot of $D_n(\theta)$:

A plot of $D_n(\theta)$,

we see that $D_n(\theta)$ oscillates with a peak at $\theta = 0$, which gets sharper with increasing n. One can prove that, see Problem 81.8,

$$\int_{-\infty}^{\infty} D_n(\theta)\, d\theta = 1 \tag{83.5}$$

and we can thus write

$$\begin{aligned} f(x) - f_n(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} (f(x) - f(y))D_n(x-y)\, dy \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} g(y)\sin((n(x-y))\, dy \end{aligned}$$

where

$$g(y) = \frac{f(x) - f(y)}{x - y}.$$

Now if $f(x)$ is differentiable with integrable derivative, it follows with an argument similar to that used in the case of Fourier series that

$$f(x) - f_n(x) =\to \ 0$$

as $n \to \infty$, which proves the following basic theorem.

Theorem 83.3 *If $f(x)$ is differentiable, then $f(x)$ is given by a convergent Fourier integral:*

$$f(x) = \int_{-\infty}^{\infty} \hat{f}(\xi)e^{i\xi x} \quad \textit{for } x \in \mathbb{R}.$$

83.4 Parsevals formula

Parseval formula takes the following form for the Fourier transform (for a proof see Problem 83.3):

Theorem 83.4 *If $f(x)$ has a convergent Fourier series representation, then*

$$\int_{-\infty}^{\infty} |f(x)|^2\, dx = 2\pi \int_{-\infty}^{\infty} |\hat{f}(\xi)|^2\, d\xi.$$

83.5 Solving the 1d heat equation on ℝ using the Fourier transform

We consider the 1d homogeneuous heat equation on $\mathbb{R}$

$$\begin{aligned} \dot{u}(x,t) - u''(x,t) &= 0 \quad \text{for } x \in \mathbb{R},\ t > 0, \\ u(x,0) &= u_0(x) \quad \text{for } x \in \mathbb{R}, \end{aligned} \tag{83.6}$$

where the initial value $u_0(x)$ is integrable over $\mathbb{R}$ and we seek a solution $u(x,t)$ which is integrable over $\mathbb{R}$ for all $t > 0$. Taking Fourier transforms with respect to x, we are led to the forllowing initial value problem for each $\xi \in \mathbb{R}$

$$\frac{d}{dt}\widehat{u}(\xi,t) + \xi^2\widehat{u}(\xi,t) = 0 \quad \text{for } t > 0, \quad \widehat{u}(\xi,0) = \widehat{u}_0(\xi)$$

with solution

$$\widehat{u}(\xi,t) = e^{-t\xi^2}\widehat{u}_0(\xi).$$

We thus obtain the following solution formula

$$u(x,t) = \frac{1}{\sqrt{2\pi t}}\int_{-\infty}^{\infty} e^{-\frac{(x-y)^2}{4t}} u_0(y)\,dy.$$

83.6 The connection between Fourier series and Fourier transforms

Suppose $f : \mathbb{R} \to \mathbb{C}$ is periodic with period $\frac{2\pi}{\omega}$ with $\omega > 0$ and has the Fourier series representation

$$f(x) = \sum_{m=-\infty}^{\infty} e^{im\omega x}, \quad c_m(f) = \frac{\omega}{2\pi}\int_{-\frac{\pi}{\omega}}^{\frac{\pi}{\omega}} f(y)e^{-im\omega y}\,dy,$$

which we write in the form

$$f(x) = \sum_{m=-\infty}^{\infty} \frac{1}{2\pi}\Big(\int_{-\frac{\pi}{\omega}}^{\frac{\pi}{\omega}} f(y)e^{-im\omega y}\,dy\Big)\, e^{im\omega x}\omega. \tag{83.7}$$

We now compare with a Fourier transform representation of a non-periodic function $f : \mathbb{R} \to \mathbb{C}$ according to the previous section:

$$f(x) = \int_{\infty}^{\infty} \frac{1}{2\pi}\Big(\int_{\infty}^{\infty} f(y)e^{-i\xi x}\,dy\Big)\, e^{i\xi x}\,d\xi. \tag{83.8}$$

We formally obtain from (83.7) by replacing $m\omega$ by ξ and ω by $d\xi$ viewing the sum over m as a Riemann sum and letting ω tend to 0.

Note the normalization used in the definition of the Fourier transform $\widehat{f}(\xi)$ with the factor $\frac{1}{2\pi}$, and the factor $\frac{\omega}{2\pi}$ in the definition of the Fourier coefficients $c_m(f)$ of a function f with period $\frac{2\pi}{\omega}$.

83.7 The sampling theorem

Let $f : \mathbb{R} \to \mathbb{C}$ be a given function with Fourier transform $\widehat{f}(\xi)$, and suppose that $\hat{f}(\xi) = 0$ for $|\xi| \geq \pi$. By Fouriers inversion formula, we have

$$f(x) = \int_{-\pi}^{\pi} \widehat{f}(\xi) e^{ix\xi}\, d\xi.$$

We now expand $\widehat{f}(\xi)$ in a Fourier series:

$$\widehat{f}(\xi) = \sum_{m=-\infty}^{\infty} c_m(\widehat{f}) e^{im\xi}$$

with Fourier coefficients

$$c_m(\widehat{f}) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \widehat{f}(\eta) e^{-im\eta}\, d\eta.$$

Using that $\widehat{f}(\eta) = 0$ for $|\eta| \geq \pi$, we can write

$$c_m(\widehat{f}) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\eta) e^{-im\eta}\, d\eta = f(-m)$$

where used Fouriers inversion formula. We thus obtain the representation formula:

$$\begin{aligned} f(x) &= \int_{-\pi}^{\pi} \sum_{m=-\infty}^{\infty} f(-m) e^{im\xi} e^{ix\xi}\, d\xi \\ &= \sum_{m=-\infty}^{\infty} f(-m) \int_{-\pi}^{\pi} e^{im\xi} e^{ix\xi}\, d\xi \\ &= \sum_{m=-\infty}^{\infty} f(-m) \frac{\sin(x+m)}{x+m} = \sum_{m=-\infty}^{\infty} f(m) \frac{\sin(x-m)}{x-m} \end{aligned}$$

which gives a representation of $f(x)$ for any value of x in terms of the values $\{f(m)\}$ with m integer. We have now prove the famous:

Theorem 83.5 (Sampling Theorem) *If $f : \mathbb{R} \to \mathbb{C}$ has a Fourier transform $\widehat{f}(\xi)$ such that $\hat{f}(\xi) = 0$ for $|\xi| \geq \pi$, then*

$$f(x) = \sum_{m=-\infty}^{\infty} f(m) \frac{\sin(x-m)}{x-m}$$

We conclude that *sampling* the values $f(m)$ of the function $f(x)$ for the integer values $m = 0, \pm 1, \pm 2, ...,$ gives information of all the values of $f(x)$ for any $x \in \mathbb{R}$, under thge assumption that $\hat{f}(\xi) = 0$ for $|\xi| \geq \pi$.

Example 83.6. The Sampling theorem takes the following form for a function $f(x)$ such that $\hat{f}(\xi) = 0$ for $|\xi| \geq a\pi$, where $a > 0$ is a constant:

$$f(x) = \sum_{m=-\infty}^{\infty} f(\frac{m}{a})\frac{\sin(ax-m)}{ax-m}.$$

This follows by applying the Sampling theorem to $g(x) = f(\frac{x}{a})$, recalling that $\widehat{g}(\xi) = a\widehat{f}(a\xi)$ and noting that $\widehat{g}(\xi) = 0$ if $|\xi| \geq \pi$ since $\hat{f}(\xi) = 0$ for $|\xi| \geq a\pi$. We see that the larger the factor a gets, that is the larger is the interval outside which the Fourier transform vanishes, the closer the sampling points $\frac{m}{a}$ will be distributed. Of course this couples to the Nyquist cut-off frequency.

83.8 The Laplace transform

We give a brief account of the *Laplace transform*, which is closely related to the Fourier transform. The Laplace transform is useful in solving certain constant-coefficient linear initial value problems analytically with classical applications in e.g. control theory.

For a given function $f : [0, \infty) \to \mathbb{R}$, we define the *Laplace transform* $Lf : [0, \infty)$ by

$$Lf(s) = \int_0^\infty e^{-st} f(t)\, dt \quad fors \in [0, \infty).$$

We denote here the independent variable by t indicating typical applications with t representing time.

Example 83.7. If $f(t) = e^{-at}$, then $Lf(s) = \frac{1}{s+a}$.

Example 83.8. If $f(t) = \frac{t^n}{n!}$ then $Lf(s) = \frac{1}{s^{n+1}}$. This follows by repeated integration by parts.

Example 83.9. If $f(t) = \sin(mt)$ then $Lf(s) = \frac{m}{m^2+s^2}$. If $f(t) = \cos(mt)$ then $Lf(s) = \frac{s}{m^2+s^2}$.

We note the following connection between the Laplace transform of $Df = f'$ and f:

$$Lf'(s) = sLf(s) - f(0) \tag{83.9}$$

which follows by integration by parts.

83.8.1 Laplace transforms and constant-coefficient linear inital value problems

The typical application goes as follows: Consider the initial value problem $u'(t) + u(t) = f(t)$ for $t > 0$ with $u(0) = 0$. Taking Laplace transforms of both sides we get

$$sLu(s) + Lu(s) = Lf(s), \quad \text{or } Lu(s) = \frac{f(s)}{s+1}$$

For example, if $f(t) = 1$, then $Lf(s) = \frac{1}{s}$ and thus $Lu(s) = \frac{1}{s(s+1)} = \frac{1}{s} - \frac{1}{s+1}$ and we conclude that $u(s) = 1 - e^{-t}$. Having a catalogue of Laplace transforms we may expect to be able to solve constant-coefficient linear inital value problems.

83.9 Wavelets and the Haar basis

We give a short introduction to wavelets in the simplest setting of 1d piecewise constant approximation using the *Haar basis*, which combines the features of orthogonality and local support. We thus consider functions defined on the unit interval $[0, 1]$ and we let $0 = x_0 < x_1 < ... < x_N = 1$ be a uniform subdivision with $x_j = jh_n$, $h_n = 2^{-n}$ and $N = 2^n$ for some natural number n. A natural orthogonal basis for the space V_n of piecewise constant functions on the subdivison $0 = x_0 < x_1 < ... < x_N = 1$ consists of the set of functions $\{\varphi_{n,k}\}_{k=0}^{N-1}$, where $\varphi_{n,k}(x) = 1$ for $x \in I_{n,k} = (kh_n, (k+1)h_n)$ and $\varphi_{n,k}(x) = 0$ else, that is, each basis function $\varphi_{n,k}(x)$ is equal to 1 on the subinterval $I_{n,k}$ and vanishes elsewhere. We can express these functions through scaling and translation of one single function in the form

$$\varphi_{n,k} = \varphi(2^n x - k) \quad \text{for ,} \quad k = 0, ..., N-1,$$

where $\varphi(x) = 1$ for $x \in (0, 1)$, and $\varphi(x) = 0$ else. We note that V_{n-1} is a subspace of V_n since the space V_n is built on a finer subdivison than V_{n-1}.

We shall now present a different orthogonal basis for V_n which displays the "difference" between V_n and V_{n-1}, and which carries useful information on the various scales in V_n. More precisely, we shall express each $u \in V_n$ in the form $u = v + w$ with $v \in V_{n-1}$ and $w \in W_{n-1}$, where W_{n-1} is spanned by the functions $\psi_{n-1,k} = \psi(2^{n-1}x - k)$ for $k = 0, ..., 2^{n-1} - 1$, expressed through scaling and translation of the single function $\psi(x)$ given by

$$\psi(x) = \begin{cases} 1 & \text{for } 0 < x < \frac{1}{2}, \\ -1 & \text{for } \frac{1}{2} < x < 1, \end{cases}$$

and $\psi(x) = 0$ else. We note that $(v, w) = \int_0^1 v(x)w(x)\,dx = 0$ if $v \in V_{n-1}$ and $w \in W_{n-1}$. Further, the two functions $\varphi_{n-1,k}$ and $\psi_{n-1,k}$ obviously span the two-dimensional space of functions on the interval $I_{n,k}$ which are piecewise constant on the two subintervals $kh_n < x < kh_n + h_{n+1}$ and $kh_n + h_{n+1} < x < kh_n + h_n$ of $I_{n,k}$. We thus have the following orthogonal decomposition

$$V_n = V_{n-1} \oplus W_{n-1},$$

stating that each function $u \in V_n$ can be expressed in the form $u = v + w$ with $v \in V_{n-1}$, $w \in W_{n-1}$ and $(v, w) = 0$, see Fig. **??**.

Illustration of the orthogonal decomposition $V_n = V_{n-1} \oplus W_{n-1}$

We can thus express V_n as an orthogonal sum:

$$V_n = V_0 \oplus W_0 \oplus W_1 \oplus ... \oplus W_{n-1}$$

where each space $\oplus W_j$ measures variations on the scale 2^{-j}. The corresponding basis functions comprise the so-called *Haar basis* for V_n combining orthogonality and local support.

Chapter 83 Problems

83.1. Solve using Fourier series the differential equation $-D^2u(x) = f(x)$ with $f(x)$ a given 2π-periodic function with zero mean value and we seek a 2π-periodic solution $u(x)$ with zero mean value.

83.2. Model the following electrical circuits: (i) resistor 1 and inductor in series over applied voltage (ii) resistor 1 and capacitor in series over applied voltage (iii) resistor 2 coupled in series with resistor 1 and inductor in parallel over applied voltage. Output voltage drop over resistor 1. Solve using Fourier series. Show that (i) and (ii) correspond to low-pass filters and (iii) to high-pass filter.

83.3. Prove Parsevals formula for the Fourier transfom. Hint: Set $\widehat{g}(\xi) = \overline{\widehat{f}}(\xi)$, which is the same as setting $g(-x) = \overline{f}(x)$, and integrate over ξ:

$$\int_{-\infty}^{\infty} f(x)\overline{f}(x)\,dx = f * g(0) = \int_{-\infty}^{\infty} \widehat{f * g}(\xi)\,d\xi = 2\pi \int_{-\infty}^{\infty} |\widehat{f}(\xi)|^2\,d\xi,$$

and use Theorem 83.2.

83.4. Prove that for $a \in \mathbb{R}$, we have (i) $\widehat{g}(\xi) = e^{-ia\xi}\widehat{f}(\xi)$ if $g(x) = f(x-a)$, (ii) $\widehat{g}(\xi) = \widehat{f}(\xi - a)$ if $g(x) = e^{-iax}f(x)$.

84
Analytic Functions Tool Bag

84.1 Differentiability and analyticity

A function $f : \Omega \to \mathbb{C}$ is *differentiable* at $z_0 \in \Omega$ with derivative $f'(z_0) \in \mathbb{C}$, if for z close to z_0, we have

$$|f(z) - f(z_0) - f'(z_0)(z - z_0)| \leq K_f(z_0)|z - z_0|^2, \qquad (84.1)$$

where $K_f(z_0)$ is a non-negative real constant depending on f and z_0.

A function $f : \Omega \to \mathbb{C}$ is *analytic* in the open domain Ω of the complex plane if $f(z)$ is differentiable at all $z_0 \in \Omega$ with derivative $f'(z_0)$. If $f : \Omega \to \mathbb{C}$ is analytic, then also $f' : \Omega \to \mathbb{C}$ is analytic with derivative $f'' : \Omega \to \mathbb{C}$, which is also analytic, and so on. An analytic function $f : \Omega \to \mathbb{C}$ thus has derivatives of all orders $f^{(n)} : \Omega \to \mathbb{C}$, $n = 1, 2, ...$, which are all analytic.

The usual rules for differentiation of sums, products and quotients valid for functions $f : \mathbb{R} \to \mathbb{R}$ extend to functions $f : \mathbb{C} \to \mathbb{C}$.

The function $f(z) = z^n$ is analytic in $\mathbb{C}$ for $n = 1, 2, ...$

The function $f(z) = z^{-n}$ is analytic for $z \neq 0$ if $n = 1, 2, ...$

84.2 The Cauchy-Riemann equations

If $f(z) = u(x, y) + iv(x, y)$ is analytic in the open domain Ω of the complex plane, then the real and imaginary parts $u(x, y)$ and $v(x, y)$ satisfy the

Cauchy-Riemann equations in Ω:

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x},$$

or in polar coordinates $z = re^{i\theta}$:

$$\frac{\partial u}{\partial r} = \frac{1}{r}\frac{\partial v}{\partial \theta}, \quad \frac{\partial v}{\partial r} = -\frac{1}{r}\frac{\partial u}{\partial \theta}.$$

84.3 The real and imaginary parts of an analytic function are harmonic

If $f : \Omega \to \mathbb{C}$ is analytic, where Ω is an open domain of the complex plane $\mathbb{C}$, then the real part $u(x,y) = \text{Re } f(z)$ and the imaginary part $v(x,y) = \text{Im } f(z)$ are harmonic in Ω.

84.4 Conjugate harmonic functions

If $u(x,y)$ is harmonic in a simply connected domain Ω in $\mathbb{R}^2$, then there exists a harmonic function $v(x,y)$, uniquely determined up to a constant, such that $f(z) = u(x,y) + iv(x,y)$ is analytic in Ω. The function $v(x,y)$ is *conjugate* to $u(x,y)$.

84.5 Curves in the complex plane

A set $\Gamma = $ Range of $\gamma = \{\gamma(t) : t \in I\}$ in an open domain Ω in the complex plane $\mathbb{C}$ parametrized by a Lipschitz continuous mapping $\gamma : I \to \Omega$, where $I = [a,b]$ is an interval of $\mathbb{R}$, is said to be a curve. The unit circle is a curve parametrized by the function $\gamma(t) = \exp(it)$ with $0 \leq t < 2\pi$. Γ is a *differentiable curve* if the corresponding parametrization $\gamma : I \to \mathbb{C}$ is differentiable on I, that is, decomposing $\gamma(t) = x(t) + iy(t)$ into real and imaginary parts,that is, if $x : I \to \mathbb{R}$ if $x(t)$ and $y(t)$ are differentiable on I.

A curve Γ with parametrization $\gamma : [a,b] \to \mathbb{C}$ is said to *closed and simple* if $\gamma(a) = \gamma(b)$ and $\gamma(s) = \gamma(t)$ only if $a = b$. A domain Ω in $\mathbb{C}$ which is bounded by a simple closed curve, is *simply connected.* A simply connected domain does not have any "holes".

84.6 An analytic function defines a conformal mapping

An analytic function $f : \Omega \to \mathbb{C}$, where Ω is an open domain in $\mathbb{C}$, is conformal in Ω in the sense that angles are preserved under the mapping $w = f(z)$.

84.7 Complex integrals

We define

$$\int_\Gamma f(z)\,dz = \int_a^b \big(u(x(t), y(t)) + iv(x(t), y(t))\big)(\dot{x}(t) + i\dot{y}(t))\,dt, \qquad (84.2)$$

where Ω is an open domain in the complex plane, Γ is a differentiable curve in $\mathbb{C}$ parametrized by $\gamma = (x, y) : [a, b] \to \mathbb{C}$, and $f = u + iv : \Gamma \to \mathbb{C}$ is Lipschitz continuous. Formally we have $dz = dx + idy = \dot{x}dt + i\dot{y}dt = (\dot{x} + i\dot{y})\,dt$.

84.8 Cauchy's theorem

If $f(z)$ is analytic in Ω and Γ is a simple closed curve in Ω enclosing a domain contained in Ω, then

$$\int_\Gamma f(z)\,dz = 0.$$

84.9 Cauchy's representation formula

If $f(z)$ is analytic in an open domain Ω, and Γ is a simple closed curve in Ω oriented counter-clockwise and enclosing the open domain Ω_Γ contained in Ω, then for $z_0 \in \Omega_\Gamma$,

$$f(z_0) = \frac{1}{2\pi i}\int_\Gamma \frac{f(z)}{z - z_0}\,dz,$$

and for $n = 1, 2, ...,$

$$f^{(n)}(z_0) = \frac{n!}{2\pi i}\int_\Gamma \frac{f(z)}{(z - z_0)^{n+1}}\,dz.$$

84.10 Taylor's formula

If $f(z)$ is analytic in a neigborhood Ω of a $z_0 \in \mathbb{C}$, then

$$f(z) = f(z_0) + f'(z_0)(z - z_0) + ... + \frac{f^{(n)}(z_0)}{n!}(z - z_0)^n + R_n(z),$$

where

$$R_n(z) = \frac{(z - z_0)^{n+1}}{2\pi i} \int_\Gamma \frac{f(\zeta)}{(\zeta - z_0)^{n+1}(\zeta - z)} \, d\zeta.$$

84.11 The Residue Theorem

If $f(z)$ is analytic in a simply connected open domain Ω, except at finitely many isolated points z_1, z_1,...,z_n in Ω, where $f(z)$ has simple or multiple poles, and Γ is a simple closed curve in Ω surrounding all the z_m counter-clockwise, then

$$\int_\Gamma f(z)\, dz = \sum_{m=1}^{n} 2\pi i \text{Res}\, f(z_m).$$

85
Fourier Analysis Tool Bag

85.1 Properties of Fourier coefficients

The Fourier coefficients $c_m(f)$ of a given 2π-periodic Lipschitz function$f : \mathbb{R} \to \mathbb{C}$ are defined by

$$c_m(f) = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)e^{-imx}\,dx \quad m = 0, \pm 1, \pm 2, ...,$$

and satisfy

$$c_m(f+g) = c_m(f) + c_m(g), \quad c_m(\alpha f) = \alpha c_m(f), \quad \text{for } \alpha \in \mathbb{C},$$
$$c_m(D^k f) = (im)^k c_m(f) \quad \text{for } k = 0, 1, 2, ...,$$

If $f : [-\pi, \pi]$ is Lipschitz continuous, then $c_m(f) \to 0$ as $|m| \to \infty$ (Riemann-Lebesgues' Lemma).

85.2 Convolution

Defining for 2π-periodic functions $f(x)$ and $g(x)$, the convolution $f * g$ by

$$(f * g)(x) = \int_{-\pi}^{\pi} f(x-y)g(y)\,dy \quad x \in \mathbb{R},$$

we have

$$c_m(f * g) = 2\pi\, c_m(f)c_m(g).$$

85.3 Fourier series representation

If $f : \mathbb{R} \to \mathbb{C}$ is 2π-periodic with piecewise Lipschitz continuous derivative, then $f(x)$ may be represented by a convergent Fourier series:

$$f(x) = \sum_{m=-\infty}^{\infty} c_m(f) e^{imx} \quad \text{for } x \in \mathbb{R}.$$

85.4 Parseval's formula

If $f(x)$ has a convergent Fourier series representation, then

$$\int_{-\pi}^{\pi} |f(x)|^2 \, dx = 2\pi \sum_{m=-\infty}^{\infty} |c_m(f)|^2.$$

85.5 Discrete Fourier Transforms

If $\{f_n\}_{n=0}^{N-1}$ is a sequence of N given complex numbers, then we may define a corresponding sequence $\{\widehat{f}_m\}_{m=0}^{N-1}$ by

$$\widehat{f}_m = \frac{1}{N} \sum_{n=0}^{N-1} f_n e^{-2\pi imn/N}, \quad \text{for } m = 0, ..., N-1,$$

and we say that the sequence $\{\widehat{f}_m\}_{m=0}^{N-1}$ is the *discrete Fourier transform* of the sequence $\{f_n\}_{n=0}^{N-1}$. We have the following inversion formula:

$$f_n = \sum_{m=0}^{N-1} \widehat{f}(m) e^{2\pi imn/N}, \quad \text{for } n = 0, ..., N-1.$$

85.6 Fourier transforms

For $f : \mathbb{R} \to \mathbb{C}$ piecewise Lipschitz continuous and integrable over $\mathbb{R}$, we define the Fourier transform of $f(x)$ for $\xi \in \mathbb{R}$ by

$$\widehat{f}(\xi) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(x) e^{-i\xi x} \, dx.$$

We have the inversion formula:

$$f(x) = \int_{-\infty}^{\infty} \widehat{f}(\xi) e^{i\xi x} \, d\xi \quad \text{for } x \in \mathbb{R},$$

under the assumption that $f(x)$ is differentiable on $\mathbb{R}$ with integrable derivative.

If $f(x) = e^{-|x|}$ for $x \in \mathbb{R}$, then

$$\widehat{f}(\xi) = \frac{1}{\pi}\frac{1}{1+\xi^2}.$$

If $f(x) = e^{-\frac{ax^2}{2}}$ for $x \in \mathbb{R}$, where $a > 0$ is a constant, then

$$\widehat{f}(\xi) = \frac{1}{2\sqrt{a}} e^{-\frac{\xi^2}{2a}}.$$

If $f(x) = 1 \quad \text{for } -a \le x \le a$ and $f(x) = 0$ else, where $a > 0$, then

$$\widehat{f}(\xi) = \frac{\sin(\xi a)}{\xi}.$$

85.7 Properties of Fourier transforms

If f and g are two functions with Fourier transforms $\widehat{f}$ and $\widehat{g}$, and $\alpha \in \mathbb{C}$, then

$$\widehat{(f+g)}(\xi) = \widehat{f}(\xi) + \widehat{g}(\xi), \quad \widehat{(\alpha f)}(\xi) = \alpha \widehat{f}(\xi).$$

If $g : \mathbb{R} \to \mathbb{R}$ be integrable and $f(x) = g(ax)$, then $\hat{f}(\xi) = \frac{1}{a}\hat{g}(\frac{\xi}{a})$.

If $f : \mathbb{R} \to \mathbb{C}$ is integrable with integrable derivative, then

$$\widehat{Df}(\xi) = i\xi \widehat{f}(\xi).$$

Defining for two integrable functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, the convolution $f * g$ by

$$(f * g)(x) = \int_{-\infty}^{\infty} f(x-y) g(y)\, dy,$$

we have

$$\widehat{f * g}(\xi) = 2\pi \widehat{f}(\xi)\widehat{g}(\xi) \quad \text{for } \xi \in \mathbb{R}.$$

Parseval formula:

$$\int_{-\infty}^{\infty} |f(x)|^2\, dx = 2\pi \int_{-\infty}^{\infty} |\widehat{f}(\xi)|^2\, d\xi.$$

85.8 The Sampling Theorem

If $f : \mathbb{R} \to \mathbb{C}$ has a Fourier transform $\widehat{f}(\xi)$ such that $\hat{f}(\xi) = 0$ for $|\xi| \ge a\pi$, where $a > 0$ is a constant, then

$$f(x) = \sum_{m=-\infty}^{\infty} f(\frac{m}{a}) \frac{\sin(ax - m)}{ax - m}.$$

86

Incompressible Navier-Stokes: quick and easy

A computable solution certainly exists. I don't care about the possible existence of uncomputable solutions. (Jack Nicolson)

Thus the methods of Lagrange and Hamilton are undoubtedly *useful* in helping us to carry out the primary task of dynamics - namely, to find out how systems move. But it would be wrong to think that this is the sole purpose of these general methods or even their main purpose. They do much more. In fact, they teach us what dynamics *really is* : It is the study of certain types of differential equations. (Synge and Griffiths, Principles of Mechanics, 1959)

My attention (was) drawn to various mechanical phenomena, for the explanation of which I discovered that a knowledge of mathematics was essential. (Reynolds)

By this research it is shown that there is one, and only one, conceivable purely mechanical system capable of accounting for all the physical evidence, as we know it in the Universe. (Reynolds)

86.1 Introduction

The Navier-Stokes equations is the basic model for fluid flow and describe a variety of phenomena in hydro and aero-dynamics, processing industry, biology, oceanography, geophysics, meteorology and astrophysics. Fluid flow in all these applicatons usually contains features of both *turbulent* and *laminar* flow, with turbulent flow being irregular with rapid fluctuations in space and time and laminar flow being more organized. The basic question of *Computational Fluid Dynamics* CFD is how to efficiently and reliably solve the Navier-Stokes equations numerically for both laminar and turbulent flow.

The Navier-Stokes equations is a system of nonlinear differential equations coupling the phenomena of convection and diffusion. Traditionally, the study of the Navier-Stokes equations is separated into *incompressible* and *compressible* flow, using different dependent variables: *primitive variables* (velocity, pressure, temperature) for incompressible flow and *conservation variables* (density, momentum, energy) for compressible flow. We focus in this chapter on the incompressible Navier-Stokes equations in the case of constant density, viscosity and temperature, with the velocity and pressure as variables. We present the cG(1)dG(0) finite element method with cG(1) in space and dG(0) in time, and follow up with the corresponding cG(1)dG(1) and cG(1)cG(1) methods.

86.2 The incompressible Navier-Stokes equations

The Navier-Stokes equations for an incompressible Newtonian fluid with constant kinematic viscosity $\nu > 0$, unit density and constant temperature enclosed in a volume Ω in $\mathbb{R}^3$ with boundary Γ, take the form: find the velocity (u, p) such that

$$\begin{array}{rcll} \frac{\partial u}{\partial t} + (u \cdot \nabla)u - \nu\Delta u + \nabla p &=& f & \text{in } \Omega \times I, \\ \nabla \cdot u &=& 0 & \text{in } \Omega \times I, \\ u &=& w & \text{on } \Gamma \times I, \\ u(\cdot, 0) &=& u^0 & \text{in } \Omega, \end{array} \tag{86.1}$$

where $u = (u_1, u_2, u_3)$ is the velocity and p the pressure of the fluid and f, w, u^0, $I = (0, T)$, is a given driving force, boundary data, initial data and time interval, respectively. Recall that

$$\frac{\partial v}{\partial t} + (u \cdot \nabla)v = \frac{\partial v}{\partial t} + \sum_{i=1}^{3} u_i \frac{\partial v}{\partial x_i} \tag{86.2}$$

is the *particle derivative* of a quantity $v(x, t)$ measuring the rate of change of $v(x(t), t)$ with respect to time, that is the rate of change of v along a trajectory $x(t)$ of a fluid particle with velocity $u(x, t)$, satisfying $\frac{dx}{dt} = u(x(t), t)$. In particular, $\frac{\partial u}{\partial t} + (u \cdot \nabla)u$ is the acceleration (rate of change of velocity) of a fluid particle. The expression $\nu\Delta u - \nabla p$ represents the total force on a fluid particle resulting from of viscous shear force and an isotropic pressure. The first equation of 86.1, which is a vector equation

$$\frac{\partial u_i}{\partial t} + (u \cdot \nabla)u_i - \nu\Delta u_i + \frac{\partial p}{\partial x_i} = f_i, \qquad i = 1, 2, 3,$$

is the *momentum equation* expressing Newton's second law stating that the acceleration is proportional to the force, and the second equation expresses the incompressibility condition. We consider here the case of Dirichlet boundary conditions with the velocity u being prescribed on the boundary Γ. Below we consider Neumann and Robin boundary conditions. Below we will often write for short $(u \cdot \nabla)u = u \cdot \nabla u$.

The linear *Stokes equations* are obtained omitting the nonlinear term $u \cdot \nabla u$, which is possible if the velocity u is small, corresponding to *creeping flow*.

The *Reynolds number Re* is defined by $Re = \frac{uL}{\nu}$, where u represents a velocity and and L a length scale characteristic of the flow. The size of the Reynolds number is decisive. If $Re \sim 1$, then the flow is very viscous, a situation met in e.g. polymer flow or forming processes. In most applications in areo/hydro-dynamics, Re is much larger than 1, often very large up to 10^6 or even larger. In these cases with small viscosity, the flow may be very complex or turbulent.

There is a stationary analog of 86.1 assuming the solution to be independent of time along with the driving force and boundary data. A stationary solution normally arises as a limit of a time-dependent solution as time tends to infinity, and this is often reflected in the computation of a stationary solution through some kind of time-stepping until convergence. For larger Reynolds numbers, stable stationary solutions in general do not not exist.

86.3 Numerical methods for the incompressible Navier-Stokes equations

Trying to solve the incompressible Navier-Stokes equations numerically, we meet the following difficulties

- instabilities from discretization of convection terms,
- pressure instabilities in equal order interpolation of velocity and pressure.

The simplest cure to convection instability is to increase the viscosity ν in the computation so that $\nu \geq uh$, where u is the local fluid velocity and h is the local mesh size. The simplest stabilization of the pressure p, is to modify the incompressibility equation $\nabla \cdot u = 0$ to $-\nabla \cdot (\delta \nabla p) + \nabla \cdot u = 0$, with $\delta \approx h^2$ with $h(x)$ the local mesh size.

In Galerkin methods the stabilization can be achieved in higher-order consistent form by adding least-squares control of residuals. We present this approach below in the context of the cG(1)dG(0) method with cG(1) in space and dG(0) in time. We also present corresponding cG(1)cG(1) and cG(1)dG(1) methods.

86.4 The cG(1)dG(0) method

We now present the cG(1)dG(0) method for 86.1 starting with the case of homogeneous Dirichlet boundary conditions. Let $0 = t_0 < t_1 < ... < t_N = T$ be a sequence of discrete time levels with associated time steps $k_n = t_n - t_{n-1}$. Let W_h be the usual finite element space of continuous piecewise linear functions on a triangulation $\mathcal{T}_h = \{K\}$ of Ω with mesh function $h(x)$. Let W_h^0 be the space of functions in W_h vanishing on Γ. We shall seek an approximate velocity $U(x,t)$ such that $U(x,t)$ is continuous and piecewise linear in x for each t, and $U(x,t)$ is piecewise constant in t for each x. Similarly, we shall seek an approximate pressure $P(x,t)$ which is continuous piecewise linear in x and piecewise constant in t. More precisely, we shall seek $U^n \in V_h^0$ with $V_h^0 = W_h^0 \times W_h^0 \times W_h^0$ and $P^n \in W_h$ for $n = 1, ..., N$, and we shall set

$$\begin{aligned} U(x,t) = U^n(x) \quad x \in \Omega, \quad t \in (t_{n-1}, t_n], \\ P(x,t) = P^n(x) \quad x \in \Omega, \quad t \in (t_{n-1}, t_n]. \end{aligned} \tag{86.3}$$

Further we write for velocites $v = (v_i)$ and $w = (w_i)$

$$(v,w) = \int_\Omega v \cdot w \, dx, \qquad (\nabla v, \nabla w) = \int_\Omega \sum_i^3 \nabla v_i \cdot \nabla w_i \, dx,$$

and similarly for scalar functions p and q defined on Ω:

$$(p,q) = \int_\Omega pq \, dx.$$

We now formulate the cG(1)dG(0) method without stabilization as follows: For $n = 1, ..., N$, find $(U^n, P^n) \in V_h^0 \times W_h$ such that

$$\begin{aligned} (\frac{U^n - U^{n-1}}{k_n}, v) + (U^n \cdot \nabla U^n + \nabla P^n, v) + (\nu \nabla U^n, \nabla v) = (f^n, v) \quad \forall v \in V_h^0, \\ (\nabla \cdot U^n, q) = 0 \quad \forall q \in W_h, \end{aligned} \tag{86.4}$$

where $U^0 = u^0$, and we set $f^n(x) = f(x, t_n)$. We see that the discrete equations result from multiplication of the momentum equation with $v \in V_h^0$ and the incompressibility equation by $q \in W_h$, followed by integration over Ω including integration by parts in the term $(-\nu \Delta U, v)$.

We can write the cG(1)dG(0) method without stabilization alternatively as follows: For $n = 1, ..., N$, find $(U^n, P^n) \in V_h^0 \times W_h$ such that

$$\begin{aligned} &(\frac{U^n - U^{n-1}}{k_n}, v) + (U^n \cdot \nabla U^n + \nabla P^n, v) + (\nabla \cdot U^n, q) \\ &+ (\nu \nabla U^n, \nabla v) = (f^n, v) \quad \forall (v,q) \in V_h^0 \times W_h, \end{aligned} \tag{86.5}$$

where we simply added the equations in 86.4.

The cG(1)dG(0) method with stabilization takes the form: For $n = 1, ..., N$, find $(U^n, P^n) \in V_h^0 \times W_h$ such that

$$(\frac{U^n - U^{n-1}}{k_n}, v) + (U^n \cdot \nabla U^n + \nabla P^n, v + \delta(U^n \cdot \nabla v + \nabla q)) + (\nabla \cdot U^n, q)$$
$$+ (\nu \nabla U^n, \nabla v) = (f^n, v + \delta(U^n \cdot \nabla v + \nabla q)) \quad \forall (v, q) \in V_h^0 \times W_h, \tag{86.6}$$

where δ is a stabilization parameter defined as follows: $\delta(x) = h^2(x)$ in the case of *diffusion-dominated* flow with $\nu \geq Uh$, and

$$\delta = (\frac{1}{k} + \frac{U}{h})^{-1} \tag{86.7}$$

in the case of *convection dominated* flow with $\nu < Uh$. Note that if $k \approx \frac{h}{U}$, which is a natural choice of time step in the convection-dominated case, then $\delta \approx \frac{1}{2}\frac{h}{U}$. Note further that the stabilized form 86.6 of the cG(1)dG(0) method is obtained by replacing v by $v + \delta(U^n \cdot \nabla v + \nabla q)$ in the terms $(U^n \cdot \nabla U^n + \nabla P^n, v)$ and (f^n, v). In principle, we should make the replacement throughtout, but in the present case of the cG(1)dG(0), only the indicated terms get involved because of the low order of the approximations. The perturbation in the stabilized method is of size δ, and thus the stabilized method has the same order as the original method (first order in h if $k \sim h$).

Letting v vary in 86.6 while choosing $q = 0$, we get the following equation (the discrete momentum equation):

$$\begin{aligned}(\frac{U^n - U^{n-1}}{k_n}, v) + (U^n \cdot \nabla U^n + \nabla P^n, v + \delta U^n \cdot \nabla v)\\ + (\nu \nabla U^n, \nabla v) = (f^n, v + \delta U^n \cdot \nabla v) \quad \forall v \in V_h^0,\end{aligned} \tag{86.8}$$

and letting q vary while setting $v = 0$, we get the following discrete pressure equation:

$$(\delta \nabla P^n, \nabla q) = -(\delta U^n \cdot \nabla U^n, \nabla q) - (\nabla \cdot U^n, q) + (\delta f^n, \nabla q) \quad \forall q \in W_h. \tag{86.9}$$

We normally seek to solve the system 86.6 iteratively alternatively solving the velocity equation 86.8 for U^n with P^n given, and the pressure equation 86.9 for P^n with U^n given.

86.5 The cG(1)cG(1) method

We present the following cG(1)cG(1) variant of the cG(1)dG(0) method with cG(1) in time instead of dG(0): For $n = 1, ..., N$, find $(U^n, P^n) \in V_h^0 \times W_h$ such that

$$
\begin{aligned}
&(\frac{U^n - U^{n-1}}{k_n}, v) + (\hat{U}^n \cdot \nabla \hat{U}^n + \nabla P^n, v + \delta(\hat{U}^n \cdot \nabla v + \nabla q)) + (\nabla \cdot \hat{U}^n, q) \\
&+ (\nu \nabla \hat{U}^n, \nabla v) = (f^n, v + \delta(\hat{U}^n \cdot \nabla v + \nabla q)) \quad \forall (v, q) \in V_h^0 \times W_h,
\end{aligned}
\tag{86.10}
$$

where $\hat{U}^n = \frac{1}{2}(U^n + U^{n-1})$. Evidently, we obtained the cG(1) version by changing from U^n to $\hat{U}^n$ in all terms but the first in the cG(1)dG(0) method.

86.6 The cG(1)dG(1) method

We shall now formulate the cG(1)dG(1) method obtained by replacing dG(0) by dG(1) in the cG(1)dG(0) method. In this method the discrete velocity $U(x,t)$ is piecewise linear linear in time on each time interval I_n, with possibly discontinuitites at the discrete time levels t_n. More precisely, we make the Ansatz:

$$
U^n(x,t) = \frac{t_n - t}{k_n} U_+^{n-1}(x) + \frac{t - t_{n-1}}{k_n} U_-^n(x), \qquad \text{for } t_{n-1} < t < t_n,
\tag{86.11}
$$

where U_+^{n-1} and U_-^n belong to V_h^0. We note that

$$
U_\pm^n(x) = \lim_{s \to 0^+} U(x, t_n \pm s)
$$

is the limit of $U(x,t)$ as t approaches t_n from below $(-)$, or above $(+)$. The cG(1)dG(1) method takes the form: For $n = 1, ..., N$, find U^n of the form 86.11 and $P^n \in W_h$, such that for all $v(x,t) = w_1(x,t) + (t - t_{n-1}) w_2(x,t)$ with $w_1, w_2 \in V_h^0$ and $q \in W_h$,

$$
\begin{aligned}
&(U_+^{n-1} - U_-^{n-1}, v) \\
&+ \int_{t_{n-1}}^{t_n} ((\dot{U}^n + U^n \cdot \nabla U^n + \nabla P^n, v + \delta(\dot{U}^n + U^n \cdot \nabla v + \nabla q)) + (\nabla \cdot U^n, q))\, dt \\
&+ \int_{t_{n-1}}^{t_n} (\nu \nabla U^n, \nabla v)\, dt = \int_{t_{n-1}}^{t_n} (f^n, v + \delta(\dot{U} + U^n \cdot \nabla v + \nabla q)).
\end{aligned}
\tag{86.12}
$$

We may similary let P be piecewise linear discontinuous in time.

86.7 Neumann boundary conditions

To properly model Neumann boundary conditions, we first need to recall that the components σ_{ij} of the *total stress tensor* $\sigma = (\sigma_{ij})$ acting on a fluid element, are given by

$$\sigma_{ij} = \bar{\sigma}_{ij} - p\delta_{ij}, \quad i,j = 1,2,3,$$

where the *stress deviatoric* $\bar{\sigma} = (\bar{\sigma}_{ij})$ is coupled to the *strain tensor* $\epsilon(u) = (\epsilon_{ij}(u))$ with components

$$\epsilon_{ij}(u) = (\partial u_i/\partial x_j + \partial u_j/\partial x_i)/2, \quad i,j = 1,2,3,$$

through the constitutive relation of a *Newtonian fluid*:

$$\bar{\sigma}_{ij} = 2\nu\epsilon_{ij}(u), \quad i,j = 1,2,3,$$

where ν is the constant viscosity, and $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$. We observe that the trace of the stress deviatoric is zero, that is,

$$\sum_{i=1}^{3} \bar{\sigma}_{ii} = 2\nu \sum_{i=1}^{3} \epsilon_{ii}(u) = 2\nu\nabla \cdot u = 0,$$

and thus the total stress σ is decomposed into a stress deviatoric $\bar{\sigma}$ with zero trace and an isotropic pressure p. Further, a direct computation shows that

$$\nu\Delta u - \nabla p = \nabla \cdot \sigma, \tag{86.13}$$

where $\nabla \cdot \sigma$ is a vector with components $(\nabla \cdot \sigma)_i$ given by

$$(\nabla \cdot \sigma)_i = \sum_{j=1}^{3} \frac{\partial \sigma_{ij}}{\partial x_j}.$$

Multiplying 86.13 by $v = (v_i)$ with $v = 0$ on Γ and integrating by parts, we find that

$$\nu(\nabla u, \nabla v) + (\nabla p, v) = 2\nu(\epsilon(u), \epsilon(v)) + (\nabla p, v),$$

where

$$(\epsilon(u), \epsilon(v)) = \sum_{i,j=1}^{3} \int_{\Omega} \epsilon_{ij}(u)\epsilon_{ij}(v)\, dx.$$

We are thus led to replace the term $(\nu\nabla u, \nabla v)$ by the term $(2\nu\epsilon(u), \epsilon(v))$ in variational formulations of the Navier-Stokes equations. In the case of Dirichlet boundary conditions for the velocity the two expressions are equal, since the test velocity v vanishes on Γ, but in the case of Neumann type boundary conditions the replacement opens the possibility of enforcing in variational form a Neuman boundary condition of the form

$$\sum_{j=1}^{3} \sigma_{ij} n_j = \sum_{j=1}^{3} \bar{\sigma}_{ij} n_j - p n_i = \sum_{j=1}^{3} 2\nu\epsilon_{ij}(u) n_j - p n_i = g_i \quad \text{on } \Gamma_2, \quad i = 1,2,3, \tag{86.14}$$

which expresses that the total force on the boundary part Γ_2 is equal to the given force $g = (g_i)$. For example, if $g = 0$, then this condition expresses that the total force is zero on Γ_2, which we may use as an outflow boundary condition simulating that the fluid freely flows out into a large reservoir. More precisely, the presence of the terms

$$-(p, \nabla \cdot v) + (2\nu\epsilon(u), \epsilon(v))$$

in a variational formulation with v varying freely on Γ_2, will enforce a homogeneous Neumann boundary condition 86.14 upon integration by parts.

We now consider a typical situation with the boundary Γ decomposed into two parts Γ_1 an Γ_2 with the velocity being equal to a given velocity w on Γ_1 and imposing the homogeneous Neumann condition 86.14 on Γ_2. For simplicity, we assume that w is independent of time, the extension to time dependence of w being evident. Typically, w will be zero on a part of Γ_1 and will be directed into Ω on the remaining part corresponding to a given inflow.

We let V_h be the space of continuous piecewise linear velocities v on a triangulation $\mathcal{T}_h = \{K\}$ of Ω with mesh function $h(x)$, satisfying the boundary condition $v = w$ on Γ_1, and let V_h^0 be the corresponding test space of functions with $v = 0$ on Γ_1. Let W_h be the space of continuous piecewise linear pressures p on $\mathcal{T}_h = \{K\}$, and W_h^0 the corresponding test space of pressures q such that $q = 0$ on Γ_2.

The stabilized cG(1)dG(0) method can be formulated as follows: For $n = 1, ..., N$ seek $U^n \in V_h$ and $P^n \in W_h$ such that

$$\begin{aligned}(\frac{U^n - U^{n-1}}{k_n}, v) + (U^n \cdot \nabla U^n, v + \delta U^n \cdot \nabla v) - (P^n, \nabla \cdot v)\\ + (2\nu\epsilon(U^n), \epsilon(v)) = (f^n, v + \delta U^n \cdot \nabla v) \quad \forall v \in V_h^0,\end{aligned} \tag{86.15}$$

$$(\delta \nabla P^n, \nabla q) = -(\delta U^n \cdot \nabla U^n, \nabla q) - (\nabla \cdot U^n, q) + (\delta f^n, \nabla q) \quad \forall q \in W_h^0, \tag{86.16}$$

where we choose P^n on Γ_2 according to 86.14 with $g = 0$ and u replaced by U. Again we seek to solve the system iteratively alternatively solving the velocity equation 86.15 for U^n with P^n given, and the pressure equation 86.16 for P^n with U^n given.

For a discussion of artificial outflow boundary conditions which may be used to truncate the domain of computation in a flow simulation, see the survey article by Rannacher.

86.8 Periodic boundary conditions

Another possibility is to truncate the computational domain by using periodic boundary conditions, where the nodes in two parts of the boundary are identified so that what goes out through one part of the boundary enters the domain through the corresponding other part of the boundary. This corresponds to the assumption that the flow in the computational domain is repeated periodically throughout the physical flow domain.

86.9 Computational examples

We now present some computational examples of 3d time dependent flows, using the stabilized cG(1)cG(1) method on a mesh with meshsize $h = 1/32$.

In Figure 86.1 we present the solution of a bluff body problem: a flow in a channel with 1x1 square cross section and length 4, with a square obstacle with side length 0.25 centered at $(0.5, 0.5, 0.5)$. We have used zero Dirichlet boundary condition for the velocity on the side walls and Neumann outflow boundary conditions on the outflow boundary. On the inflow a parabolic velocity is prescribed.

In Figure 86.2 we present the solution of a step down problem in a similar channel with a step down of height and length 0.5.

Finally in Figure 86.3 we present computations of transition to turbulence in a circular jet flow with streamwise velocity 1 in the jet and zero outside the jet, where we apply a small random perturbation. Here we have used periodic boundary conditions in all directions.

For more details on these computations see [**?**] and [**?**].

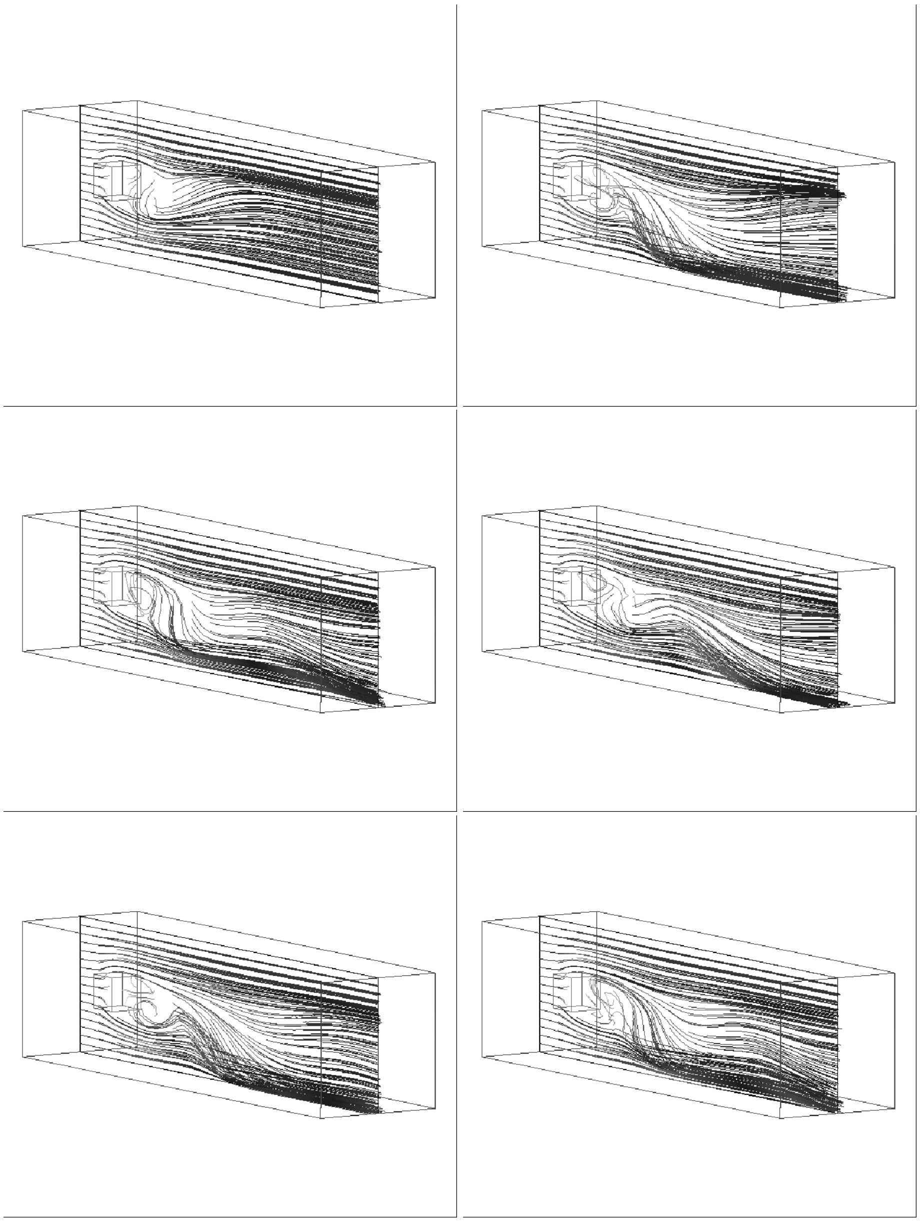


FIGURE 86.1. Bluff body flow computations for $t = 2, 4, 6, 8, 10, 12$.

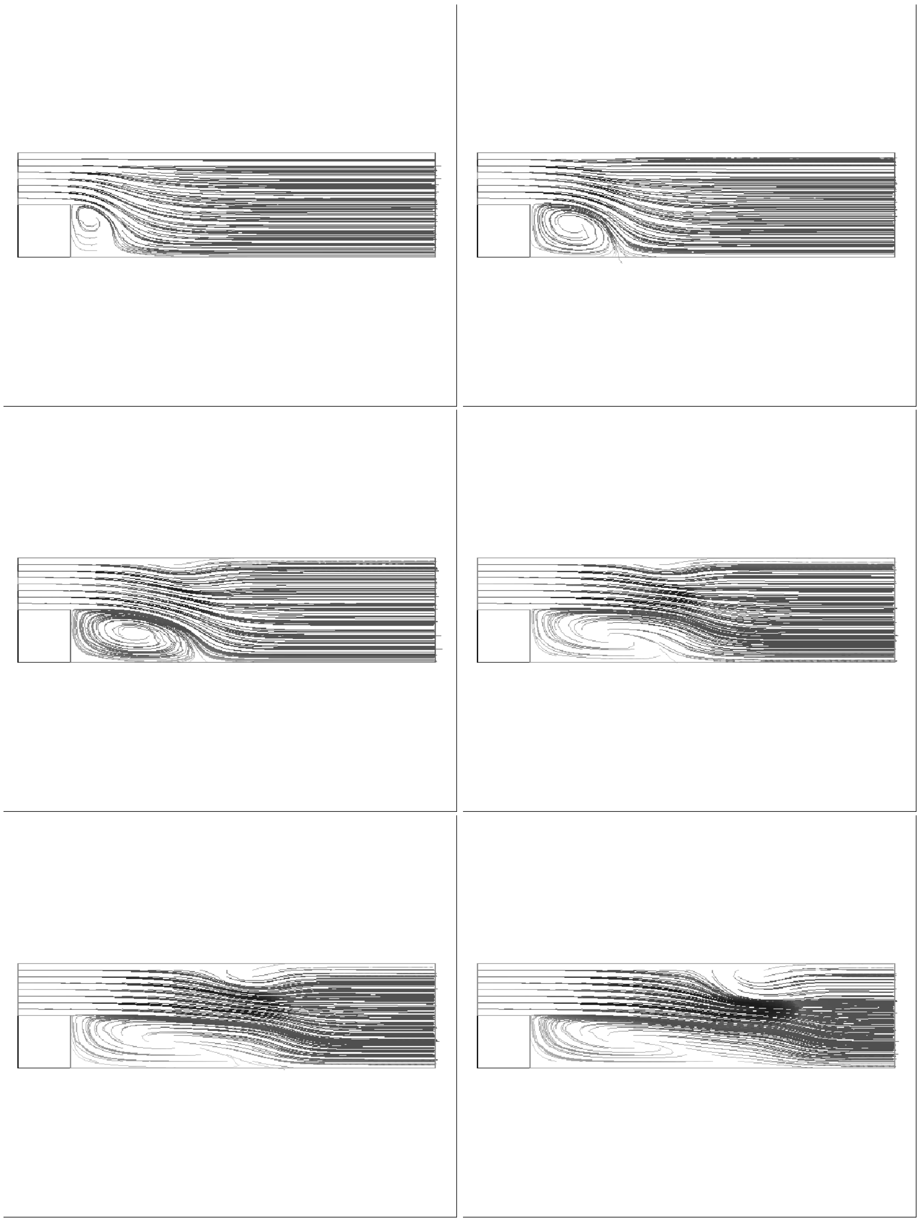


FIGURE 86.2. Step down flow computations for $t = 1, 2, 3, 4, 5, 6$.

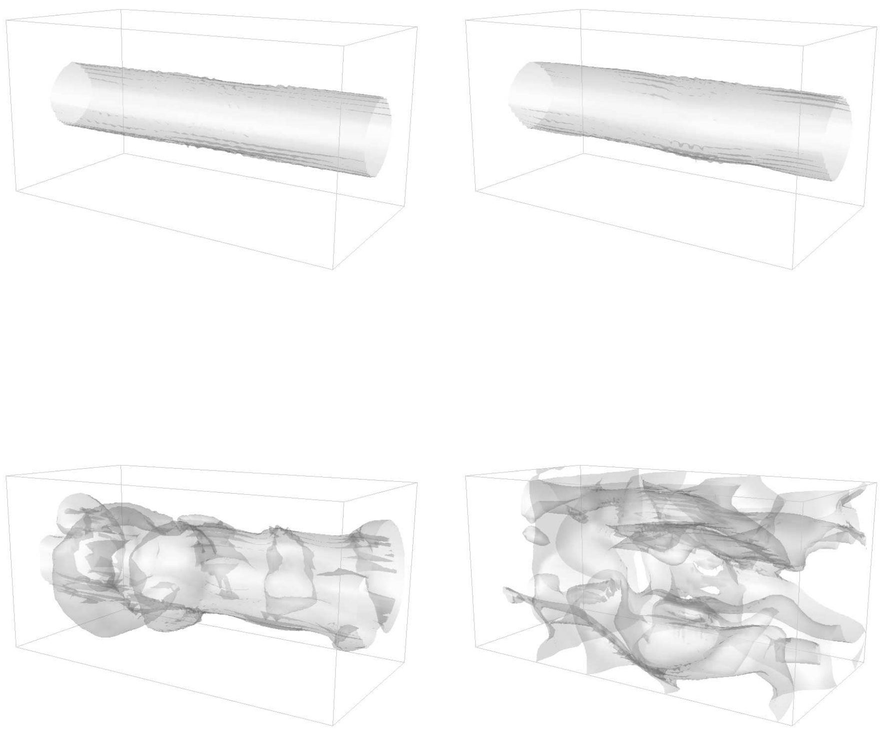


FIGURE 86.3. Streamwise velocity isosurfaces for $|u_1| = 0.02$ in jet flow for $t = 5, 7, 10, 15$

Index

K, 1028
L_2 projection, 752
V_h, 1031
$\mathbb{N}$, 54
$\mathbb{Q}$, 83
$\mathbb{R}$, 209
δ_z, 1004
$\epsilon - N$ definition of a limit, 178
λ_i, 1033
φ_i, 1034
τ_K, 1028
h refinement, 1029
h_K, 1028
$\mathcal{N}_h$, 1028
$\mathcal{S}_h$, 1028
$\mathcal{T}_h$, 1028

a posteriori error estimate, 678, 771, 773, 826, 1054
a priori error estimate, 671, 678, 770, 772
Abacus, 2
absolutely convergent series, 562
acceleration, 424
action integral, 713
adaptive
    algorithm, 773, 1055
    error control, 773, 1054
adaptive error control, 823
advancing front, 1029
alternating series, 562
Ampere's law, 957, 999
analytic geometry, 281
angle, 334
arc length, 893
Archimedes principle, 972
Archimedes' principle, 967
arclength, 899
area, 457, 915
    of triangle, 307
automatic step-size control, 824
automatized computation, 1
automatized production, 1

Babbage, 2
Babbage-Scheutz Difference Machine, 2
backward Euler, 590
bandwidth, 665
barbers paradox, 240
basis, 314, 618
bisection algorithm, 198, 227
block structure, 1048
Bolzano, 228

Bolzano's theorem, 228
boundary condition
  Dirichlet, 985
  essential, 768, 1060
  natural, 768, 1060
  Neumann, 985
  Robin, 769, 985
boundary value problem, 761
buoyancy force, 973
butterfly effect, 841

Cantor, 244
capacitor, 731
Cauchy sequence, 202, 208
Cauchy's inequality, 608
  weighted, 771
Cauchy-Riemann equations, 1091
Cauchy-Schwarz inequality, 608
center of mass, 967
centripetal acceleration, 890
change of basis, 642
change of order of integration, 910
change of variable, 935
chaos, 841
charge, 731, 999, 1087
children, 1029
Cholesky's method, 665
circle of curvature, 901
compact factorization, 665
complex number, 365
complex plane, 365
computational cost, 232
computer representation of numbers, 188
condition number, 670
conservation
  heat, 984
conservation of energy, 707
conservation of mass, 989
conservative system, 707
constitutive equation, 985
constitutive law, 989
constructivist, 241
continuous Galerkin cG(1), 590
continuous Galerkin method cG(1), 824
contour plot, 807
contraction mapping, 261
Contraction mapping theorem, 798
coordinate system, 283
Coriolis
  acceleration, 890
  force, 887
Coulomb's law, 999
Cramer's formula, 633
crash model, 724
current, 731
curvature, 900
curve, 782
curve integral, 893

dashpot, 716
de Moivres formula, 528
deca-section algorithm, 203
decimal expansions
  non-periodic, 78
  periodic, 78
delta function, 1004
derivative, 379
  of x^n, 385
  chain rule, 400
  computation of, 389
  definition, 382
  inverse function, 407
  linear combination rule, 398
  one-sided, 403
  quotient rule, 401
Descartes, 104
determinant, 627
diagonal matrix, 658
diagonally dominant, 684
diameter of a triangle, 1028
dielectric constant, 999, 1087
difference quotient, 388
differentiability
  uniform, 392
differentiable, 391, 786, 1020
differentiation under the integral sign, 803
dinner Soup model, 27
directional derivative, 795
divergence, 879, 880
divergence theorem, 944, 951
domain, 108
double integral, 905

Einstein, 433
Einstein's law of motion, 433

elastic membrane, 990
elastic string, 737
electric circuit, 731
electric conductivity, 999, 1087
electric current, 999, 1087
electric displacement, 999, 1087
electric field, 999, 1000, 1087
electrical circuit, 731
electrostatics, 1000
element, 1028
    basis function, 1033
    stiffness matrix, 1049
elementary functions, 531
elementary row operations, 658
elliptic, 988
energy, 1051
energy norm, 1051
Engdahl, 397
ENIAC, 2
equidistribution of error, 497, 774, 1058
error representation formula, 747, 773, 1036, 1053
essential boundary condition, 768, 1060
Euclid, 89
Euclidean norm, 292
Euclidean plane, 103, 283
Euler equations, 998
Eulerian description, 993
existence of minimum point, 868

Faraday's law, 999
fill-in, 667
five-point formula, 992
fixed point, 259
floating point arithmetic, 188
force field, 897
formalist school, 237
forward Euler, 590
Fourier, 429
Fourier's law, 430, 985
Fredholm, 1007
Fredholm integral equation, 1007
front, 1029
function, 107
    a^x, 511
    polynomial, 123
function $y = x^r$, 255
functions
    combinations of, 147
    rational, 149
    several variables, 169
fundamental solution, 1005
Fundamental Theorem of Calculus, 442, 454
fundamental theorem of linear algebra, 641

Galileo, 424
Gauss, 104
Gauss transformation, 660
Gauss' theorem, 941, 944, 949, 951
Gauss-Seidel method, 677
geometrically orthogonal, 297
global basis function, 1033
global stiffness matrix, 1049
GPS, 9, 97
GPS navigator, 286
gradient, 789, 880, 883
gradient field, 898
Gram-Schmidt process, 639
gravitational field, 1003
greatest lower bound, 874
Green's formula, 941, 944, 949, 951
Gulf Stream, 891
Gustafsson, Lars, 207

hanging chain, 524
hat function, 749
heat
    capacity coefficient, 984
    conduction, 983
    conductivity, 985
    flux, 984
    source, 984
heat equation, 986
Hilbert, 1007
Hooke, 427
Hooke's law, 427

identity matrix, 323
ill-conditioned matrix, 670
implicit differentiation, 408
Implicit Function theorem, 802
implicit function theorem, 809, 811
income tax, 376
incompressible, 995
independent variable, 109

induction, 734
  mutual, 734
inductor, 731
infinite decimal expansion, 207, 594
initial value problem
  general, 583
  scalar autonomous, 567
  second order, 589
  separable scalar, 575
integer, 49
  computer representation of, 61
  division with remainder, 59
integral
  additivity over subintervals, 464
  change of variables, 469
  linearity, 466
  monotonicity, 467
integral equation, 1007
integration by parts, 471, 944
interior minimum point, 869
intermediate value theorem, 228
intuitionist, 241
invariance
  orthogonal transformation, 358
inverse
  of matrix, 355
Inverse Function theorem, 801
inverse matrix, 635
irrotational, 964
isobars, 887
isotropic, 883
isotropy, 1028
iterated integration, 932
iterated one-dimensional integration, 910
iteration matrix, 674
iterative method, 667

Jacobi method, 677
Jacobian, 786, 1020
Jacquard, 2

Kirchhoff's laws, 733
Kronecker, 244

Lagrange, 701
Lagrangian description, 994
Laplace, 1003
Laplacian, 879, 881, 883
  polar coordinates, 881, 1022
  spherical coordinates, 885
LCR-circuit, 731
least squares method, 643
Leibniz, 108, 442
Leibniz' teen-age dream, 43
level curve, 669, 807
level surface, 811
liars paradox, 240
limit, 185
  computation of, 185
line, 341
line integral, 893, 898
linear combination, 294, 609
linear convergence, 671
linear function, 621
linear independence, 314, 611, 689
linear mapping, 316
linear oscillator, 717
  damped, 718
linear transformation, 357, 621
linearization, 789
linearly independent, 356
Lipschitz continuity, 155, 217
  boundedness, 165
  composition of functions, 168
  generalization, 257
  linear combinations, 163
  linear function, 156
  monomials, 163
  product of functions, 166
  quotient of functions, 166
Lipschitz continuous, 785, 1020
Lipschitz continuous function
  converging sequence, 183
load vector, 765, 1043, 1048
logarithm, 483
logicists, 238
logistic equation, 570
long division, 79
Lorenz, 841
Lorenz system, 842
lower triangular matrix, 658
lumped mass quadrature, 1038

magnetic field, 885, 999, 1087
magnetic flux, 999, 1087
magnetic permeability, 999, 1087
magnetostatics, 1002

Malthus, 432
marginal cost, 376
mass conservation, 994
mass-spring system, 703
mass-spring-dashpot systems, 715
matrix, 318, 352, 622
  factorization, 659
  ill-conditioned, 670
  multiplication, 623, 689
matrix addition, 320
matrix multiplication, 321
Maxwell, 999
Maxwell's equations, 999
Mean Value theorem, 791
medical tomography, 9
mesh, 749
  isotropy, 1028
mesh function, 497, 749, 1028
mesh modification criterion, 774
minimization method, 668
minimization problem, 668, 1051
minimum point, 866
minimum value, 866
model
  crash, 724
  infection, 581
  marriage crisis, 728
  national economy, 581
  population, 728
  spread of infection, 728
  stock market, 728
  symbiosis, 728
  transition to turbulence, 728
moment of inertia, 928, 938
muddy yard model, 30
multi-grid method, 1050
multiplication by scalar, 610

N-body system, 711
natural boundary condition, 768, 1060
natural logarithm, 483
natural number, 49
Navier-Stokes equations, 998
navigator, 285
Newton, 977
Newton's Inverse Square Law, 977
Newton's law of gravitation, 1005
Newton's Law of motion, 424
Newton's method, 413, 802
nightmare, 977
nodal basis function, 749
non-Euclidean geometry, 104
non-periodic decimal expansion, 208
norm, 292
  energy, 1052
norm of a symmetric matrix, 626, 654
numerical quadrature, 490

Ohm's law, 999
optimal mesh, 1055
optimization, 865
ordered n-tuples, 606, 688
ordered pair, 288
orthogonal, 333
orthogonal complement, 638
orthogonal decomposition, 298, 637
orthogonal matrix, 357, 639
orthogonal projection, 752
orthogonalization, 639

parallel
  lines, 311
parallelogram law, 289
parametrization, 783
parents, 1029
partial derivative, 410
partial derivatives of second order, 795
partial fractions, 537
partial pivoting, 664
particle paths, 995
partition, 749
Peano axiom system, 243
pendulum
  double, 707
  fixed support, 704
  moving support, 705
periodic decimal expansion, 208
permutation, 627
pivoting, 663
plane, 342
point mass, 1004
Poisson's equation, 987, 1042, 1057
  minimization problem, 1051
  variational formulation, 1042
Poisson's equation on a square, 1044
polar coordinates, 918

polar representation, 293
polynomial, 123
  coefficients, 123
positive series, 559
positive-definite, 766
potential, 898
potential field, 963
potential flow, 996
precision, 1038
prime number, 60
principle of equidistribution, 1055
principle of least action, 701
projection, 298, 319, 334
  point onto a line, 311
  point onto a plane, 346
projection onto a subspace, 635
Pythagoras, 89
Pythagoras' theorem, 89

QR-decomposition, 640
quadrature, 443
  adaptive, 496
  endpoint rule, 494
  lumped mass, 1038
  midpoint rule, 494
  trapezoidal rule, 494
quadrature error, 492
quarternions, 366

radius of curvature, 901
range, 108
rate of change, 375
rate of convergence, 671
rational number, 73
Reagan, 941
real number, 209
  absolute value, 212
  addition, 209
  Cauchy sequence, 215, 594
  comparison, 213
  division, 212
  multiplication, 212
reference triangle, 1046
refinement strategy, 1055
residual error, 678, 773, 1053
resistor, 731
Riemann sum, 916, 934
rigid transformations, 882
Robin boundary conditions, 769
rocket propulsion, 430
rotation, 302, 879, 881

scalar product, 333, 607
search direction, 668
separable scalar initial value problem, 575
sequence, 173
  limit of, 173
series, 558
Slide Rule, 2
socket wrench, 175
solid of revolution, 937
sorting, 866
space capsule, 973
sparse matrix, 667, 766
sparsity pattern of a matrix, 1048
spectral radius, 675
spectral theorem for symmetric matrices, 650
spherical coordinates, 884, 935
splitting a matrix, 673
square domain, 1044
squareroot of two, 195
stability
  of motion, 709
stability factor, 823
stability of floating body, 973
standard basis, 610
steepest ascent, 793
steepest descent, 793, 872
steepest descent method, 669
step length, 668
stiffness matrix, 765, 1043, 1047
Stoke's theorem, 955
Stokes, 956
Stokes theorem, 960
stopping criterion, 419, 774
straight line, 308
streamlines, 995
string theory, 737
strong boundary condition, 768
subtractive cancellation, 681
support, 1034
surface, 784
surface area, 923
surface integral, 923, 927, 1024
surface of revolution, 926
surveyor, 285

Svensson's formula, 992, 1083
symmetric matrix, 766
system linear equations, 349
system of linear equations, 312

tangent plane, 789
Taylor's theorem, 475, 798
temperature, 984
tent functions, 1034
test of linear independence, 632
total energy, 1051
transpose, 625
transpose of a matrix, 322
triangular domain, 1065
triangulation, 1028
    boundary nodes, 1057
    internal nodes, 1028, 1057
trigonometric functions, 516
triple integral, 931
triple product, 339
Turing, 236
two-body, 708
two-point boundary value problem, 761

union jack triangulation, 1065
upper triangular matrix, 658

variable, 108
variation of constants, 544
Vasa, 967
vector, 287
vector addition, 289
vector product, 304, 335
vector space $\mathbb{R}^n$, 606
Verhulst, 570
voltage, 731
Volterra-Lotka's predator-prey model, 578
volume, 933
    parallelepiped, 338
volume under graph, 912

weather prediction, 12
weight, 771
weighted L_2 norm, 771
weighted Cauchy's inequality, 771
Winnie-the-Pooh, 173

zero pivoting, 663