Applied Mathematics: Body & Soul
Vol I–III

Kenneth Eriksson, Don Estep and Claes Johnson

July 29, 2003

To the students of Chemical Engineering at Chalmers during 1998–2002, who enthusiastically participated in the development of the reform project behind this book.

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 millennium. We share this conviction with a rapidly increasing number of researchers and teachers of both 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 those 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 financial markets.

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 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 drives 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 at 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 undergraduate 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 complementary material with step-by step instructions for self-study, problems with solutions, and projects, all freely available on-line from the web site of the book. Our ambition is 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 I–III* from first year for the students of chemical engineering at Chalmers since the Fall 99, and we have used parts of the material from *AM IV–* in advanced undergraduate/beginning graduate 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 the start based on constructive mathematical methods including a computational methodology for differential equations.
- The program contains, as an integrated part, software at different levels of complexity.
- The student acquires solid skills of implementing computational methods and developing applications and software using Matlab.
- The synthesis of mathematics and computation opens 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 dynamics, as well as corresponding multiphysics 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 at 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 (from the perspective of traditional education) to both 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.

Thus the idea is 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 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 already lead the student in the second quarter 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 overall 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 continuous 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 which build on mathematical modeling.

New possibilies are opened by creating from the start a synthesis of constructive and symbolic mathematics representing a synthesis of Body & Soul: 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 and broader and lift it to a higher level. This idea underlies 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 honors 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 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 a real-valued function $f(x)$ is Lipschitz continuous with Lipschitz constant L_f on an 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, 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 seeking constructive processes for generating Cauchy sequences.

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 give a good deal of attention to the fundamental aspect of the constructibility of real numbers (viewed as possibly never-ending 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 much too 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 a line of thought leading to some result, and then we state a corresponding theorem as a summary of the hypothesis and the main result obtained. 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 hypotheses, 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 a proof is being presented; it will just be 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 hypotheses. 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 preceding 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 and basic calculus for functions of several variables; other pieces are supposed to be used as models for the students own computer realizations; 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 deep down into constructive computation, which also gives a direct coupling to applications and physical reality.

Go to http://www.phi.chalmers.se/bodysoul/

The *Applied Mathematics: Body & Soul* project has a web site containing additional instructional material and the *Mathematics Laboratory.* We hope that the web site for the student will be a good friend helping to (independently) digest and progress through the material, and that for the teacher it may offer inspiration. We also hope the web site may serve as a forum for exchange of ideas and experience related the project, and we therefore invite both students and teachers to submit material.

Acknowledgment

The authors of this book want to thank sincerely the following colleagues 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 of our reform project. And again, sincere thanks to all the students of chemical engineering at Chalmers who carried the burden of being exposed to new material often in incomplete form, and who have given much enthusiastic criticism and feed-back.

The source of mathematicians pictures is the MacTutor History of Mathematics archive, and some images are copied from old volumes of Deadalus, the yearly report from The Swedish Museum of Technology.

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 Volume 1

Contents Volume 2

Contents Volume 3

Volume 1

Derivatives and Geometry in $\mathbb{R}^3$

$$|u(x_j) - u(x_{j-1})| \leq L_u|x_j - x_{j-1}|$$

$$u(x_j) - u(x_{j-1}) \approx u'(x_{j-1})(x_j - x_{j-1})$$

$$a \cdot b = a_1b_1 + a_2b_2 + a_3b_3$$

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-

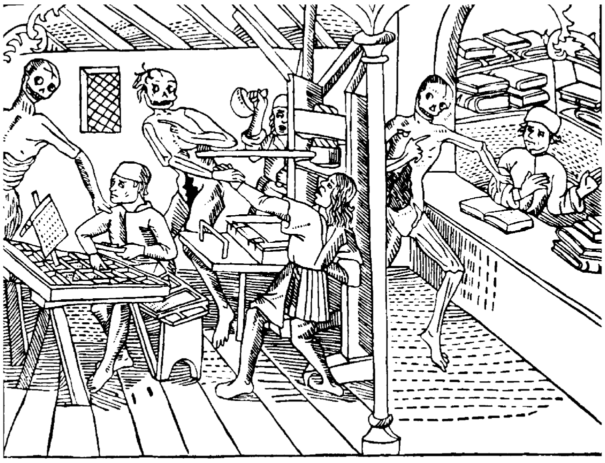


Fig. 1.1. First picture of book printing technique (from Danse Macabre, Lyon 1499)

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 weaving loom (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 were 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 square feet 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. A modern laptop at a cost of $2000 with a processor speed of 2 GHz and internal mem-

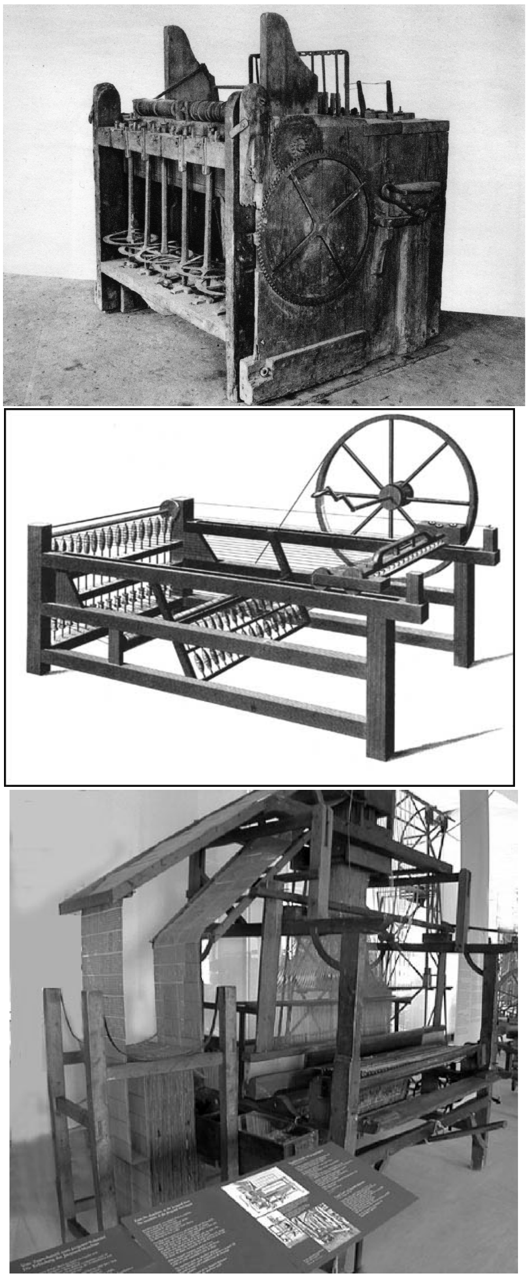


Fig. 1.2. Christoffer Polhem's machine for clock gears (1700), Spinning Jenny (1764) and Jaquard's programmable loom (1801)

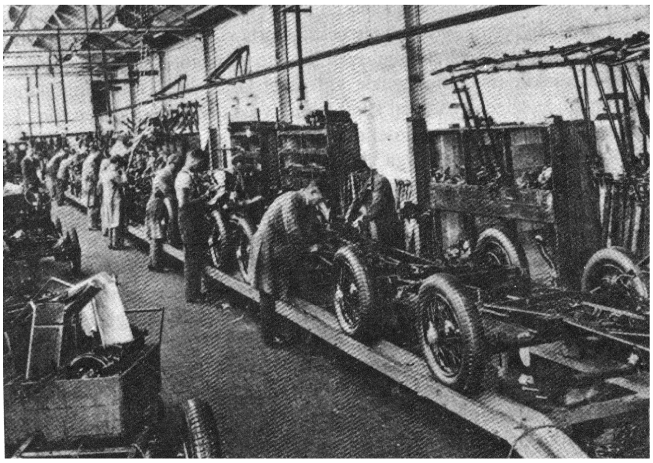


Fig. 1.3. Ford assembly line (1913)

ory of 512 Mb has the computational power of hundreds of thousands of ENIACs.

Automatization (or automation) is based on frequent repetition 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 → atom → molecule and molecule → cell → organism → population, or the patterns of our society: individual → group → society or computer → computer network → 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 mod-*

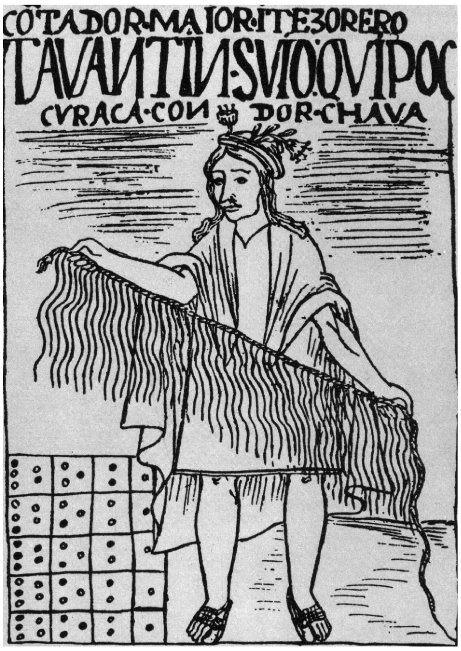


Fig. 1.4. Computing device of the Inca Culture

els in the form of *differential equations.* Using the models, real phenomena 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.

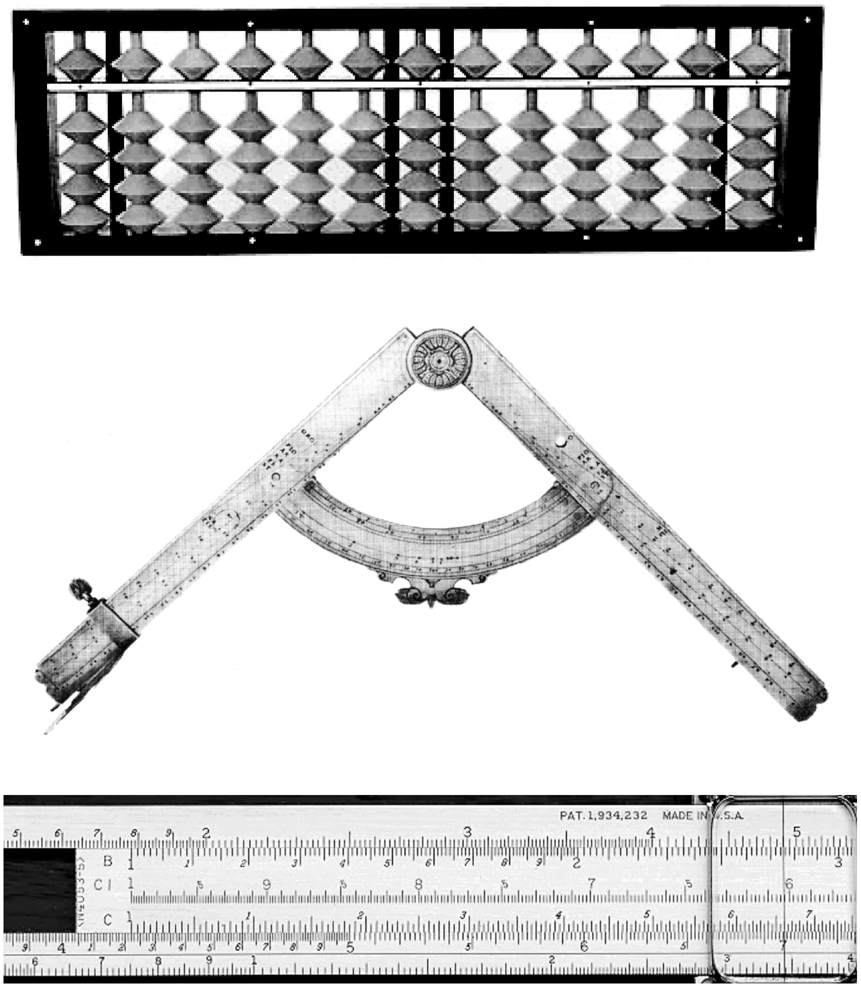


Fig. 1.5. Classical computational tools: Abacus (300 B.C.-), Galileo's Compass (1597) and Slide Rule (1620-)

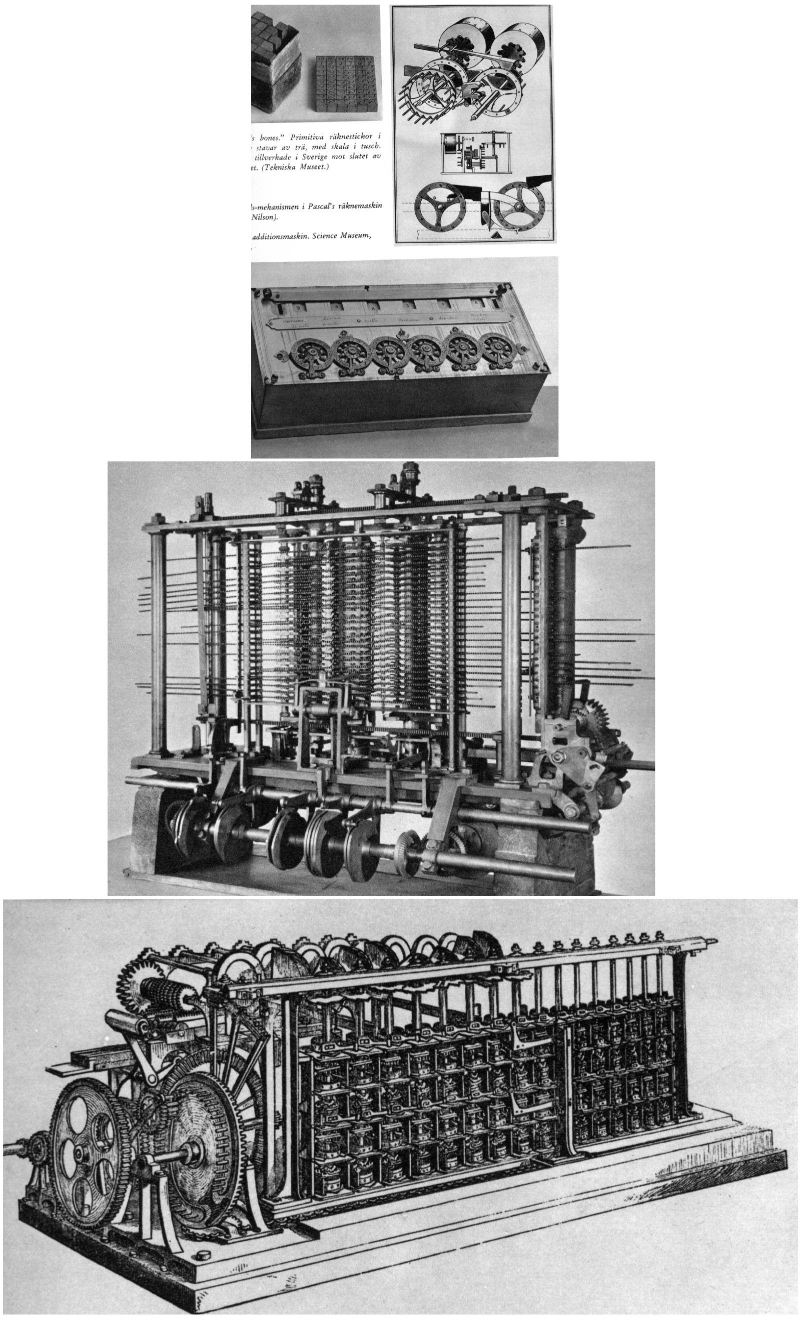
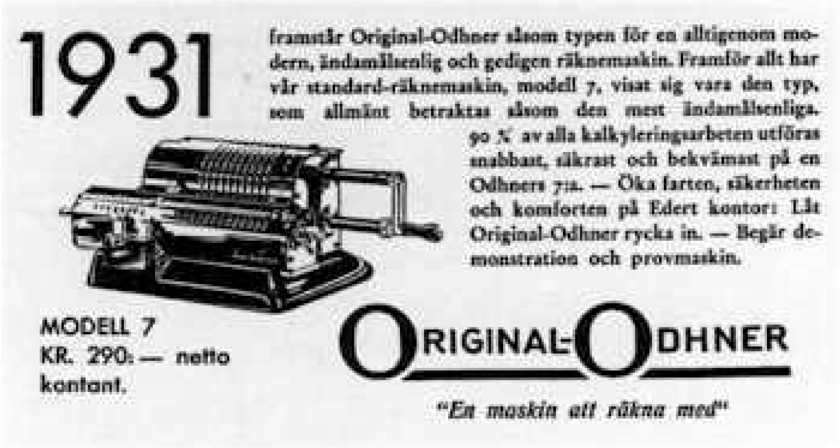

's bones." Primitiva räknestickor i
stavar av trä, med skala i tusch.
tillverkade i Sverige mot slutet av
et. (Tekniska Museet.)
ls-mekanismen i Pascal's räknemaskin
Nilson).
additionsmaskin. Science Museum,

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



1931 framstår Original-Odhner såsom typen för en alltigenom modern, ändamålsenlig och gedigen räknemaskin. Framför allt har vår standard-räknemaskin, modell 7, visat sig vara den typ, som allmänt betraktas såsom den mest ändamålsenliga. 90 % av alla kalkyleringsarbeten utföras snabbast, säkrast och bekvämast på en Odhners 7:a. — Öka farten, säkerheten och komforten på Edert kontor: Låt Original-Odhner rycka in. — Begär demonstration och provmaskin.

MODELL 7
KR. 290:— netto
kontant.

ORIGINAL-ODHNER

"En maskin att räkna med"

Fig. 1.7. Odhner's mechanical calculator made in Göteborg, Sweden, 1919–1950

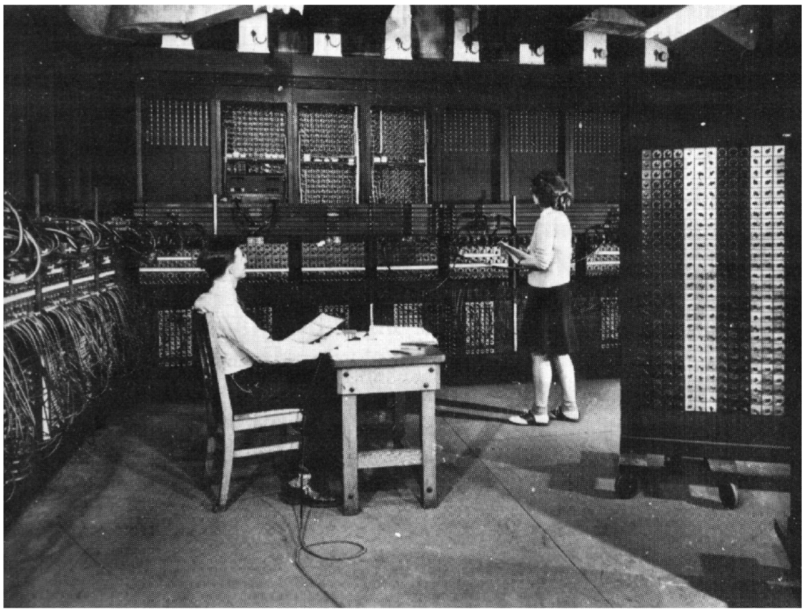


Fig. 1.8. ENIAC Electronic Numerical Integrator and Calculator (1945)

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 geometry 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

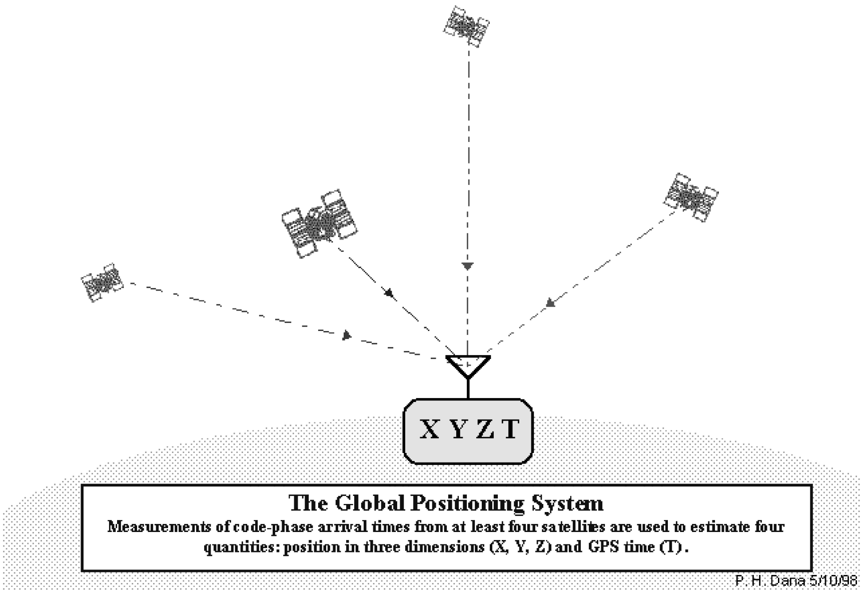

X Y Z T
The Global Positioning System
Measurements of code-phase arrival times from at least four satellites are used to estimate four quantities: position in three dimensions (X, Y, Z) and GPS time (T).
P. H. Dana 5/10/98

Fig. 1.9. GPS-system with 4 satellites

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.

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

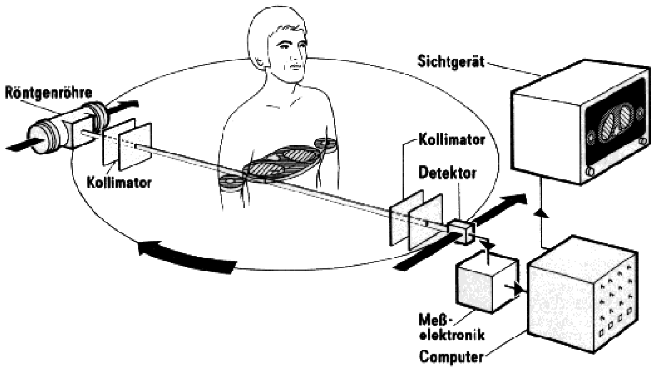


Fig. 1.10. Medical tomograph

molecules and provide likely ranges of interactions, thus greatly reducing the amount of laboratory testing time that is needed.

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.

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.

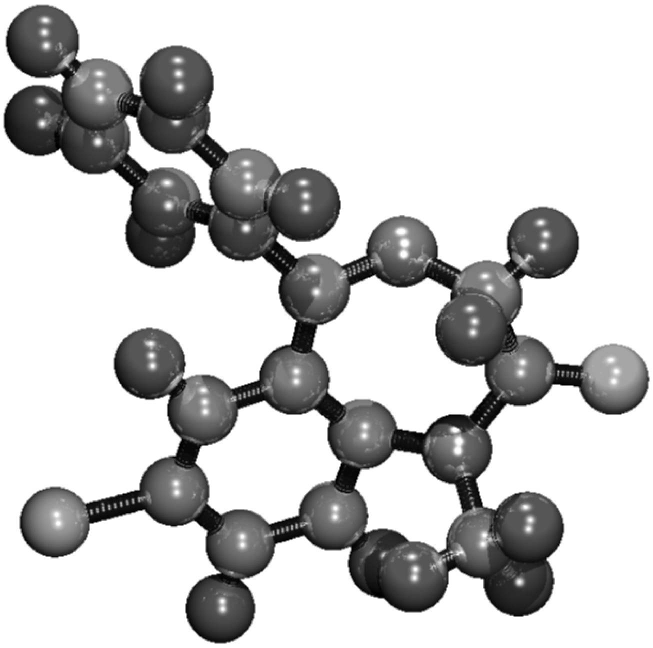


Fig. 1.11. The Valium molecule

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 equation, Laplace's equation, Navier's equation, Navier-Stokes' equations, Boussinesq's equation, Einstein's equation, Schrödinger's equation, Black-Scholes formula..., most 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 Viewing 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 advisable 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 Mathematics 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.

ASSAR
av Ulf Lundkvist
”ÄR DET INTE KONSTIGT ATT MAN KAN RÄKNA UT HUR LÅNGT DET ÄR TILL SIRIUS ALDEBARAN OCH CASSIOPEJA?
JAG TYCKER DET ÄR KONSTIGARE ATT MAN KAN VETA VAD DOM HETER

Fig. 1.12. 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

It is nothing short of a miracle that modern methods of instruction
have not yet entirely strangled the holy curiosity of inquiry.
(Einstein)

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* and *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 variables.

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 software like *DOLFIN* Dynamic Object Oriented Library for FInite Elements, and *Tanganyika* for multi-adaptive solution of systems of ordinary differential equations.

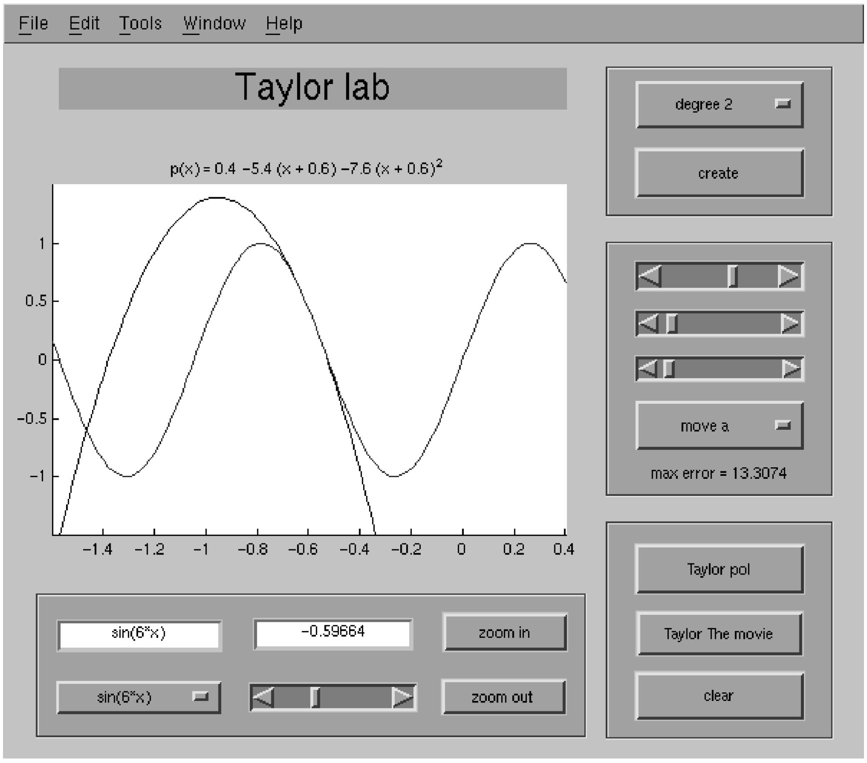
Students Lab contains constributions from students project work and will hopefully serve as a source of inspiration transferring know how from old to new students.

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.

The following labs are available from the book web page:

- Func lab – about relations and functions, inverse function etc.
- Graph Gallery – elementary 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 Newton's method
- Taylor lab – polynomials
- Opti lab – elementary optimization
- Piecewise polynomial lab – about piecewise polynomial approximation
- Integration lab – Euler and Riemann summation, adaptive integration



$p(x) = 0.4 \ -5.4\ (x + 0.6) \ -7.6\ (x + 0.6)^2$

degree 2

sin(6*x)

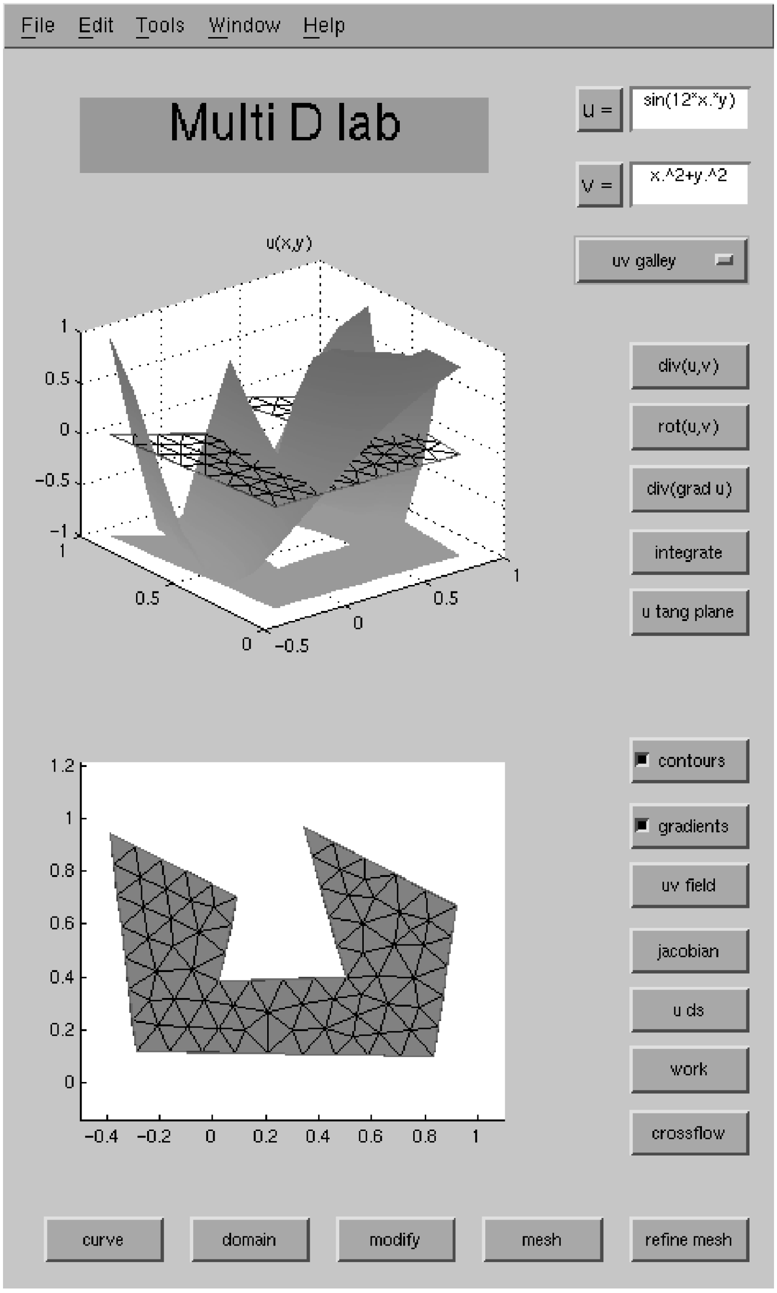
−0.59664

sin(6*x)

max error = 13.3074

Fig. 2.1. The Taylor lab

- 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



u = sin(12*x.*y)

v = x.^2+y.^2

Fig. 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 $x = \frac{2}{3} = 0.66666\ldots$. 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 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?

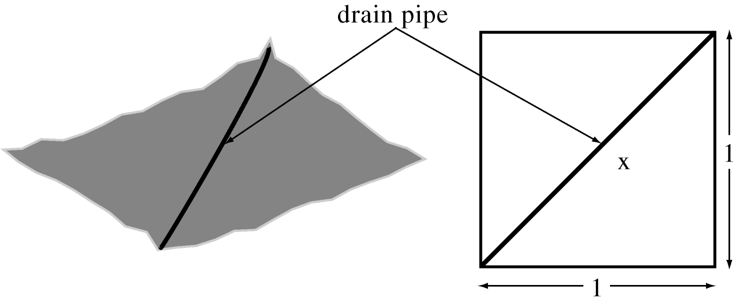


Fig. 3.1. Perspective of a field with poor drainage and a model describing the dimensions

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: The 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 the volumes of this book. 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* and *Calculus Tool Bag II*, which support the idea that a distilled 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.

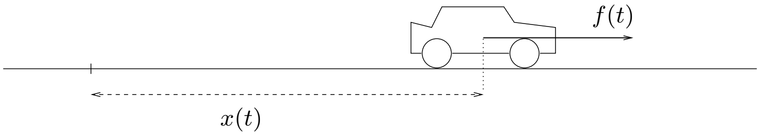


Fig. 4.1. Highway with car (Volvo?) with velocity $f(t)$ and travelled distance $x(t)$

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, $t_2, \ldots$, with corresponding distances $x(t_1)$, $x(t_2), \ldots$, 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.

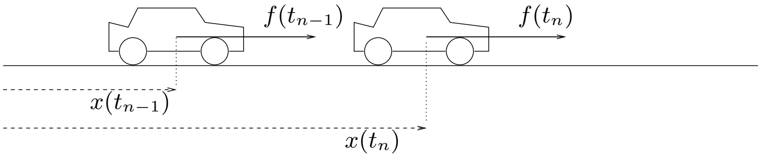


Fig. 4.2. Distance and velocity at times t_{n-1} and t_n

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 (and the veolicity does not change very quickly).

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 (4.9):

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

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 $n = 1, 2, \ldots, N$ choosing $k = 0.05$, and plot the resulting values of X_n approximating $x(t_n)$ in Fig. 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

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

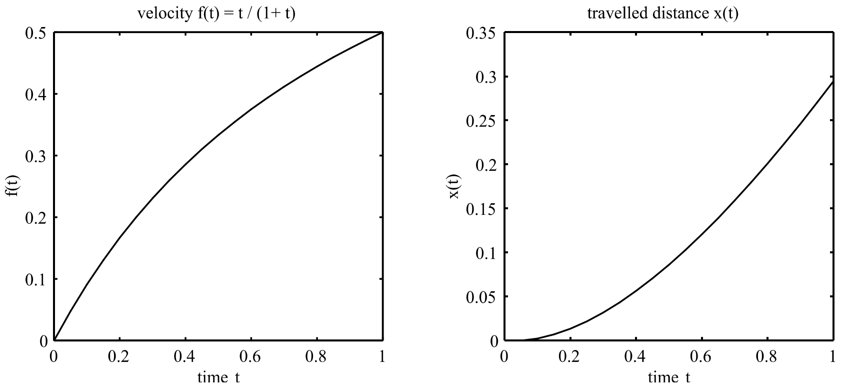


Fig. 4.3. Travelled distance X_n approximating $x(t)$ for $f(t) = \frac{t}{1+t}$ with time steps $k = 0.05$

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 $n = 1, \ldots, N$. 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

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

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

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

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 $n = 1, 2, \ldots, N$, 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 $n = 1, \ldots, N$, 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 $n = 1, \ldots, N$, we get the formula

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

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

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

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*:

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

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 $e \approx 2.72$.

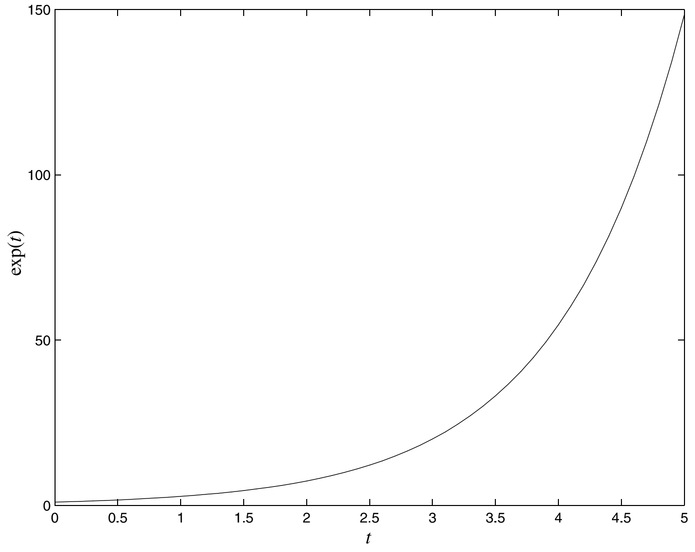


Fig. 4.4. Graph of exp(t): Exponential growth

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.

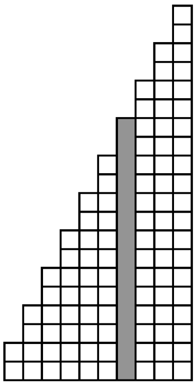


Fig. 4.5.

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

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

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.

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

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

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 Genius

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.

Here is a description of Leibniz from Encyclopedia Britannica: “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 civilization".

Fig. 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 $r = 1, 3, \frac{1}{2}$.

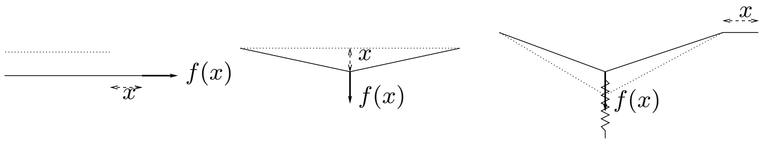


Fig. 4.7.

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.

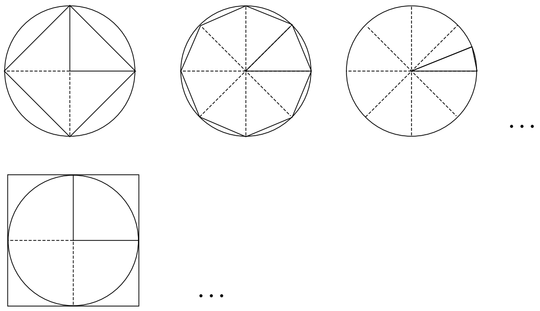


Fig. 4.8.

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.

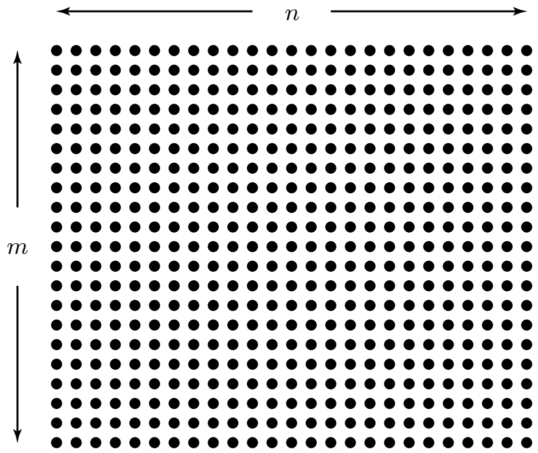


Fig. 5.1. Illustration of the commutative rule for multiplication $m \times n = n \times m$. We get the same sum if first add up the dots by counting across the rows or down the columns

In a similar way we can prove the *associative rule for multiplication*

$$m \times (n \times p) = (m \times n) \times p \tag{5.2}$$

and the *distributive rule* combining addition and multiplication,

$$m \times (n + p) = m \times n + m \times p, \tag{5.3}$$

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, $2+3\times 4$ means $2+(3\times 4) = 24$, not $(2+3)\times 4 = 20$. To overrule this convention we may use parentheses, as in $(2+3)\times 4 = 5\times 4$. From (5.3) (and (5.1)) we obtain the useful formula

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

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

$$n^p = \begin{array}{c} n \times n \times \cdots \times n \\ (\,p \text{ factors}) \end{array}$$

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

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

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.

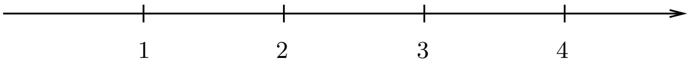


Fig. 5.2. The natural number line

We can interpret all of the arithmetic operations using the number 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 $n^0 = 1$.

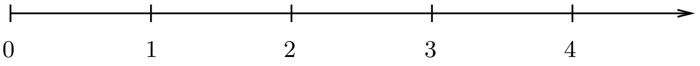


Fig. 5.3. The extended natural number line, including 0

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 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 $0, 1, 2, \ldots, 9$, 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 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 $\{1, 2, \ldots, 100\}$ of the first 100 natural numbers $1, 2, 3, \ldots, 99, 100$ 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 high rise, 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.

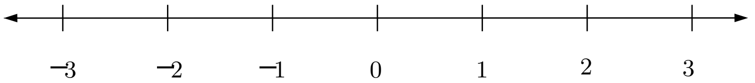


Fig. 5.4. The integer number line

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 $\{-1, -2, \ldots\}$ from the given natural numbers $\{1, 2, \ldots\}$ 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

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

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 $m, 2m, 3m, \ldots$, and noting that there must be a unique p such that $pm \leq n < (p+1)m$, see Fig. 5.5.

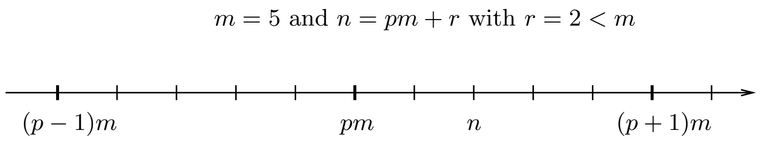


Fig. 5.5. Illustration of $pm \leq n < (p+1)m$

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

$$
\begin{aligned}
63 &= 15 + 48\\
63 = 15 + 15 + 34 &= 2 \times 15 + 33\\
63 &= 3 \times 15 + 18\\
63 &= 4 \times 15 + 3,
\end{aligned}
$$

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.TS[a]

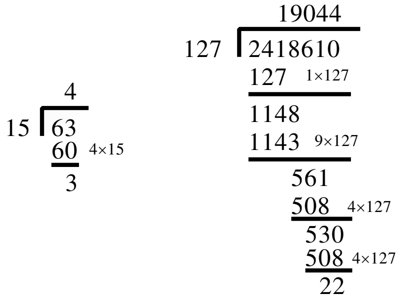


Fig. 5.6. Two examples of long division

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

TS[a] Please confirm this reference to Fig. 5.5.

Editor's or typesetter's annotations (will be removed before the final TEX run)

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 $\{-32768, -32767, \ldots, 32767\}$, which are the numbers that take two bytes of storage, and a long integer variable is an integer in the range $\{-2147483648,\ -2147483647, \ldots,$ $2147483647\}$, 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:TS[b]

TS[b] Is there a figure missing here, or should the colon be changed to dot?

Editor's or typesetter's annotations (will be removed before the final TEX run)

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) $\left(n^p\right)^q = n^{pq}$ 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 $n + 1$.

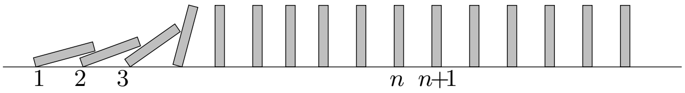


Fig. 6.1. The principle of induction: If n falls then $n + 1$ will fall, so if 1 falls they all fall!

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 the following formula for the sum of a *finite geometric series*:

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

where p is the *quotient*, which we here assume to be a given natural number, and n is a natural number. Note that we think of p as being fixed and the induction is on the number n with $n + 1$ being the number of terms in the series. 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

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

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

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

Fig. 6.2. Gauss 1831: “I protest against the use of infinite magnitude as something completed, which in mathematics is never permissible. Infinity is merely a *facon parler*TS[a], the real meaning being a limit which certain ratios approach indefinitely near, while others are permitted to increase without restriction”

TS[a] Do you mean “*façon de parler*”?

Editor’s or typesetter’s annotations (will be removed before the final TEX run)

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 × 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, $p_{m-1}, \dots, p_0$, $q_0, \dots, q_n$ 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.

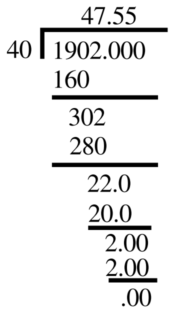


Fig. 7.1. Using long division to obtain a decimal expansion

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:

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

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 continues without ever stopping. We can find many examples of infinite decimal

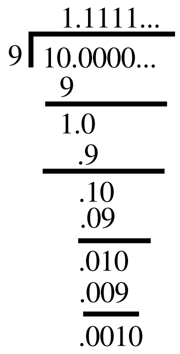


Fig. 7.2. The decimal expansion of 10/9 never stops

expansions:

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

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:

$$\begin{aligned}\frac{1043}{439} = 2.&3758542141230068337129840546697038724373576309 7\\ &9498861047835990888382687927107061503416856492 0\\ &2733485193621867881548974943052391799544419134 3\\ &9635535307517084282460136674259681093394077448 7\\ &4715261958997722095671981776765 \quad 37585421412300\\ &6833712984054669703872437357630979498861047835 9\\ &9088838268792710706150341685649202733485193621 8\\ &6788154897494305239179954441913439635535307517 0\\ &8428246013667425968109339407744874715261958997 7\\ &22095671981776765 \cdots\end{aligned}$$

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

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

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.

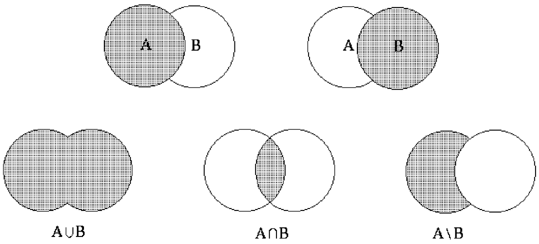


Fig. 7.3. The sets $A \cup B$, $A \cap B$ and $A \backslash B$

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.

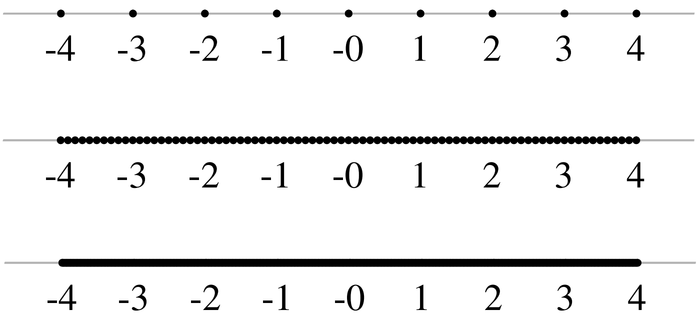


Fig. 7.4. Filling in the rational number line between -4 and 4 starting with integers, rationals with one digit, and rationals with two digits, and so on

We see that there are quickly so many points to plot that the 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

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

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

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

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

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.

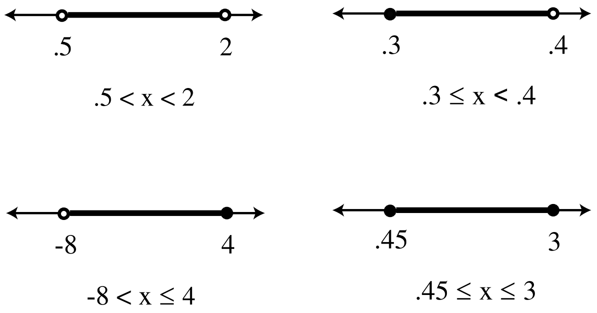


Fig. 7.5. Various rational line intervals

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 $n \ge 1$. 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:

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

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

$$\mathrm{A}_x\mathrm{B}_y \rightleftharpoons x\,\mathrm{A}^{y+} + y\,\mathrm{B}^{x-} \tag{7.11}$$

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

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

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 $\mathrm{Ba(IO_3)_2}$ in a .020 mole/liter solution of $\mathrm{KIO_3}$:

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

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

$$[\mathrm{IO_3^-}] = (.02 + 2S).$$

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

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

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

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

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

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) $\{x \in \mathbb{Q} : -1 < x \leq 2 \text{ and } 0 < x < 4\}$

(c) $\{x \in \mathbb{Q} : -1 \leq x \leq 3 \text{ or } -2 < x < 2\}$

(d) $\{x \in \mathbb{Q} : x \leq 1 \text{ or } x > 2\}$.

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 heros; Pythagoras and Euclid. 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 length c, then (see Fig. 8.1).

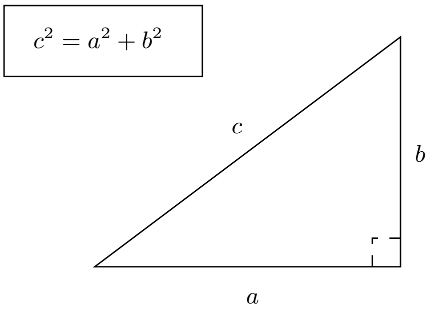


Fig. 8.1. Pythagoras' theorem

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. 8.2. 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. 8.2. This is because $\alpha + \beta = 90°$, 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.

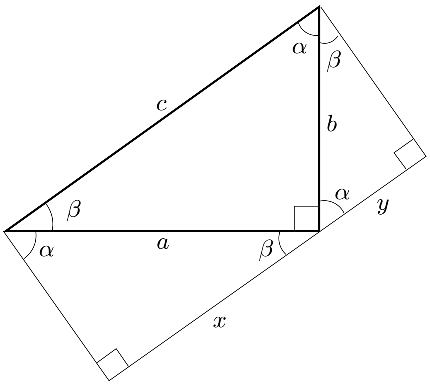


Fig. 8.2. Proof of Pythagoras theorem

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. 8.3. We recall that the angle between two lines (or line 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.

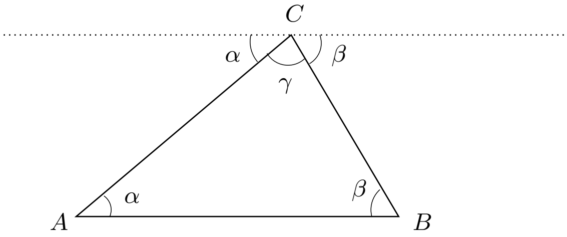


Fig. 8.3. The angles of a triangle sum up to 180°

Returning to the triangle in Fig. 8.3, 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. 8.4. This statement is Euclid's famous fifth *axiom* 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. 8.5. In the axioms, undefined concepts like *point* and *straight line* appear. When we think of these concepts, we use our intuition from

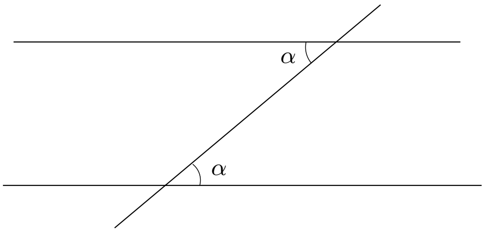


Fig. 8.4. Two parallel lines intersected by a third one

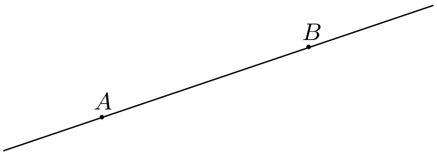


Fig. 8.5. Euclids first axiom: through two different points there is a unique straight line

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 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. 8.6, 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°.

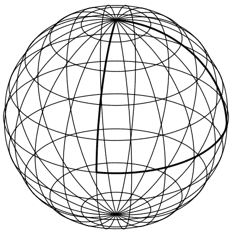


Fig. 8.6. A "triangle" on the Earth with one corner at the North Pole and two corners at the Equator with longitude difference of 90°

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 8.4.

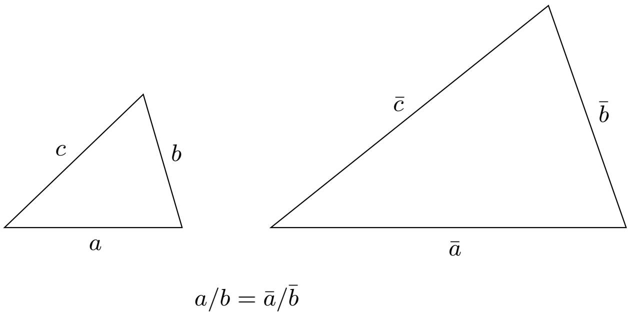


Fig. 8.7. The sides of similar triangles are proportional

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. 8.8. 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 lines passes through a point A and the other through a point B in the plane, see Fig. 8.8. 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. 8.9 We could try to check this directly with a portable right angle, but that would always leave some room for incorrect decision if the dimen-

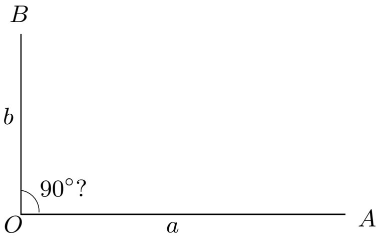


Fig. 8.8. Are these two (pieces of) lines orthogonal?

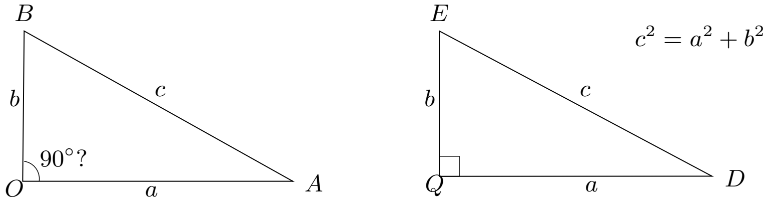


Fig. 8.9. Test of right-angledness

sion 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. 8.9, then the angle BOA would be 90°. We 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 $a = 3$, $b = 4$ 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. 8.10. We conclude that

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

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.

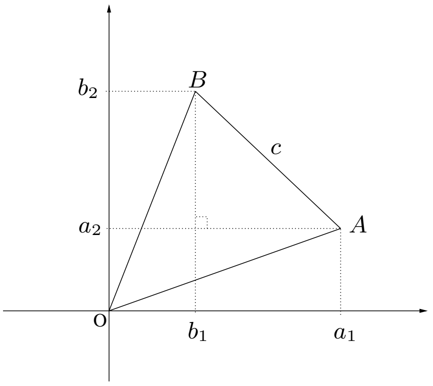


Fig. 8.10.

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!

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. 8.11. Note that Euclid's third axiom tells us that the two 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 cor-

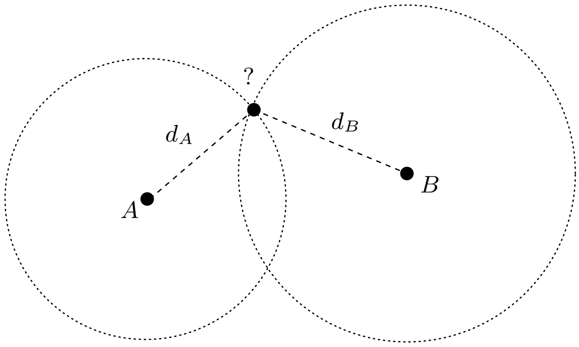


Fig. 8.11. Positioning using GPS

responding 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 over-determined 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. 8.12. 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 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. 8.12 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. 8.13, 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 $90° \leq v < 180°$.

8.8 Geometric Proof of Addition Formulas

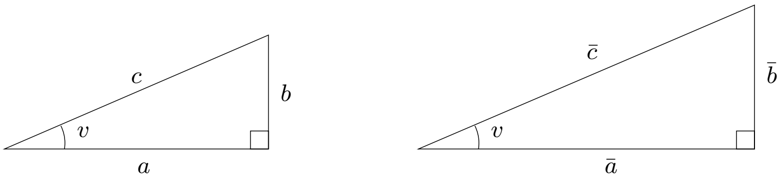


Fig. 8.12. $\cos(v) = a/c$ and $\sin(v) = b/c$

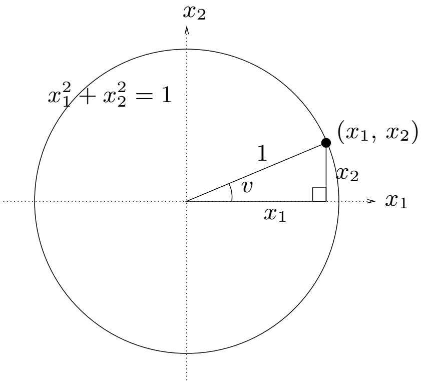


Fig. 8.13. $\cos(v) = x_1$ and $\sin(v) = x_2$

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,

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

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. 8.14 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. 8.14, 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 $\cos(\beta - \alpha) = \cos(\beta)\cos(\alpha) + \sin(\beta)\sin(\alpha)$.

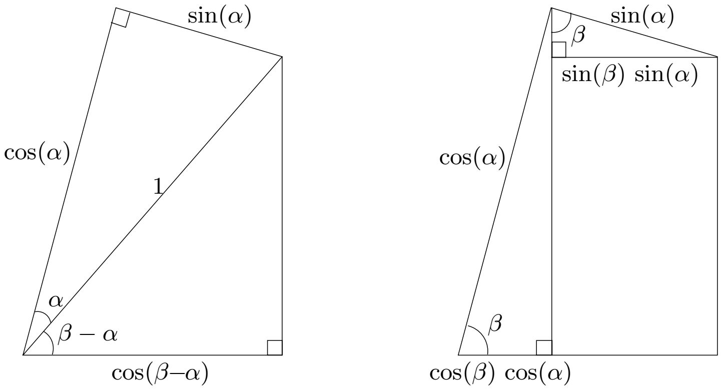


Fig. 8.14. Why $\cos(\beta - \alpha) = \cos(\beta)\cos(\alpha) + \sin(\beta)\sin(\alpha)$

From Fig. 8.14, we can similarly conclude that

$$\sin(\beta - \alpha) = \sin(\beta)\cos(\alpha) - \cos(\beta)\sin(\alpha). \tag{8.5}$$

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. 8.13, 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. 8.15. Below we will come back to these basic formulas and motivate them more closely.

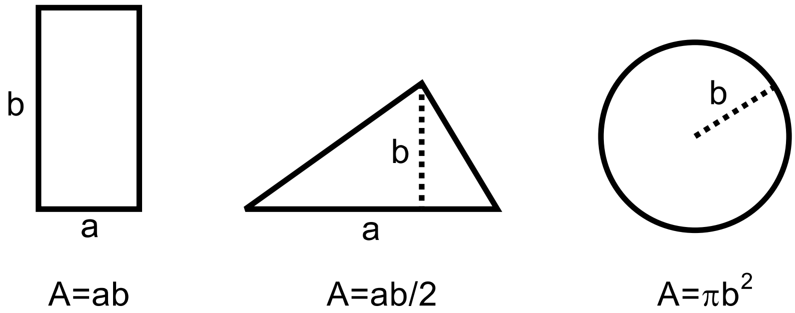


Fig. 8.15. Some common formulas for areas A

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. 8.16.

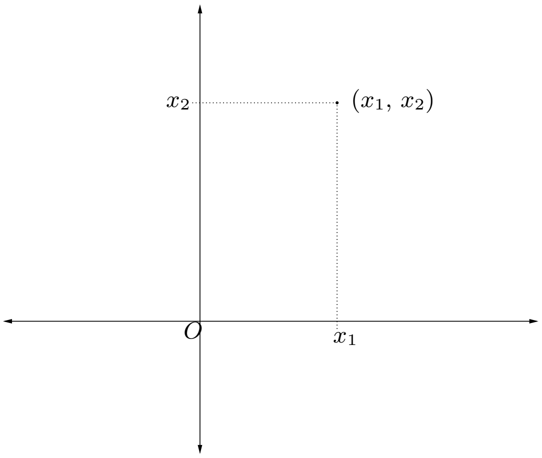


Fig. 8.16. Coordinate system of the Euclidean plane

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 (latitude-longitude) $(57.25, 12.60)$? Determine your own coordinates.

8.2. Derive (8.5) from Fig. 8.14.

8.3. Give another proof of Pythagoras' Theorem.

8.4. (a) Prove from Euclid's axioms that the sides of two triangles with the same angles are proportional (Hint: use the 5th axiom a lot). (b) 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). (c) 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.

Fig. 8.17. Two classical tools

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 $x \rightarrow f(x)$ to signify that x is mapped onto $f(x)$. We also think of the function f as a "machine" that transforms x into $f(x)$:

$$x \overset{f}{\rightarrow} f(x),$$

see also Fig. 9.1.

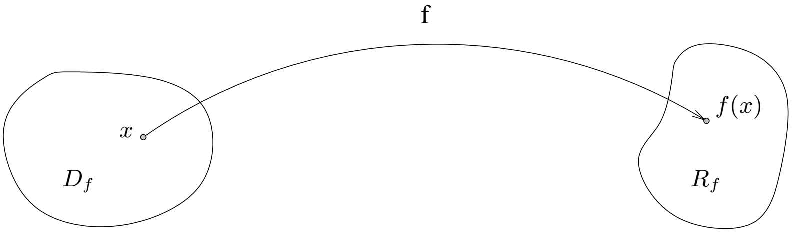


Fig. 9.1. Illustration of $f : D_f \rightarrow R_f$

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 $f : D(f) \rightarrow R(f)$ 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 $x \rightarrow f(x) = 15x$ 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 $R(f) = \{0, \pm 1, \pm 2, \pm 4, \ldots\}$. We illustrate in Fig. 9.2.

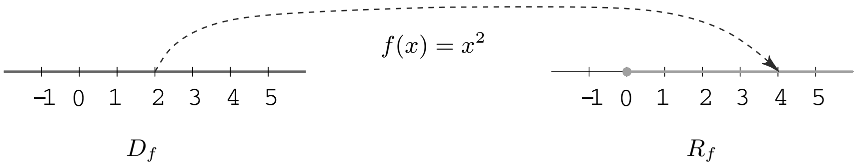


Fig. 9.2. Illustration of $f : \mathbb{Q} \to \mathbb{Q}^+$ with $f(x) = x^2$

Example 9.2. For the function $f(z) = z + 3$ we may choose, for example, $D(f) = \mathbb{N}$ and $R(f) = \{4, 5, 6, \ldots\}$, 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 $R(f) = \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots\}$.

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 $D(f) = \{1, 2, \ldots, 365\}$ 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 $1, 2, \dots 16$ 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 cetera) was entered in the form of holes in certain positions on a dollar bill 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 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

Fig. 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”

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 phone book 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 while the intersection point of the two number lines is called the *origin* and is denoted by 0.

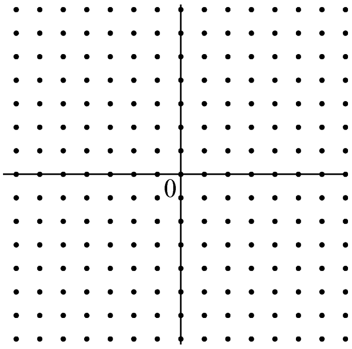


Fig. 9.4. The integer coordinate plane

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. 9.5.

Example 9.11. In Fig. 9.6, we plot n, n^2, and 2^n along the vertical axis with $n = 1, 2, 3, \ldots, 6$ along the horizontal axis. The plot suggests 2^n grows more quickly than both n and n^2 as n increases. In Fig. 9.7, 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. 9.6.

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

n	f(n)
0	0
1	1
-1	1
2	4
-2	4
3	9
-3	9
4	16
-4	16
5	25
-5	25
6	36
-6	36

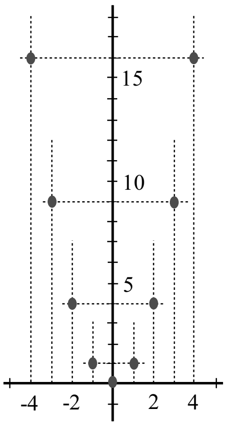


Fig. 9.5. A tabular listing of $f(n) = n^2$ and a graph of the points associated with the function $f(n) = n^2$ with domain equal to the integers

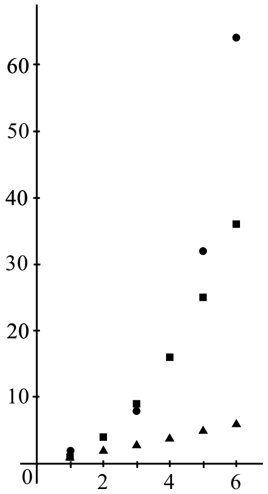


Fig. 9.6. Plots of the functions ▲ $f(n) = n$, ■$f(n) = n^2$, and ● $f(n) = 2^n$ with $D(f) = \mathbb{N}$

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

$$\{(0,0),\ (1,1),\ (-1,1),\ (2,4),\ (-2,4),\ (3,9),\ (-3,9),\ \cdots\}.$$

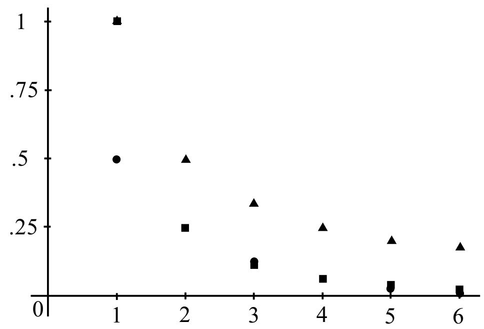


Fig. 9.7. Plots of the functions ▲ $f(n) = n^{-1}$, ■ $f(n) = n^{-2}$, and ● $f(n) = 2^{-n}$ with $D(f) = \mathbb{N}$

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. 9.5. 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 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. 9.8.

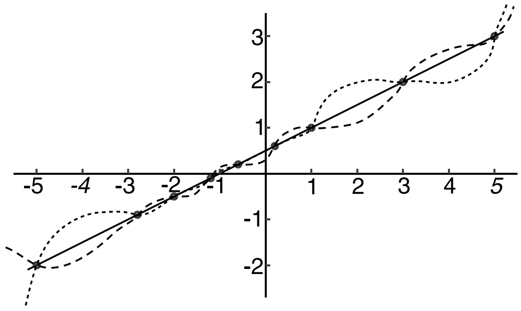


Fig. 9.8. A plot of the function values of $f(x) = \frac{1}{2}x + \frac{1}{2}$, and several functions taking on the same values at the sample points

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. 9.8. Therefore, to graph a function accurately, we would need to evaluate it at many more points than we have used in Fig. 9.8 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
-4	16
-3.5	12.25
-3.1	9.618
-2	4
-1.8	3.24
-1.4	1.96
-1	1
$-.8$	.64
$-.4$	.16
0	0
.2	.04
1.2	1.44
1.5	2.25
2.21	4.8841
2.3	5.29
2.4	5.76
3	9
3.1	9.61
3.6	12.96
3.7	13.69
4	16

and then plot the function values in Fig. 9.9.

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

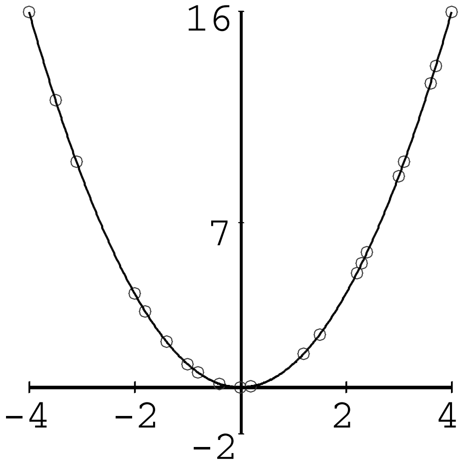


Fig. 9.9. A plot of some of the points given by $f(x) = x^2$ and a smooth curve that passes through the points

$f(x, y) = 15x + 3y$ 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 $(x, y) \to f(x, y) = 15x+3y$, denoting the pair of x and y by (x, y).

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 $f:\mathbb{N}\to\mathbb{Q}$ using two intervals where $f(n)=2^{-n}$.

9.11. Plot the following functions $f:\mathbb{Z}\to\mathbb{Z}$ 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 $f:\mathbb{Q}\to\mathbb{Q}$ 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

$$f(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots + a_nx^n, \qquad (10.1)$$

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 $f(x) = a_0 + a_1x$ of degree 1 and *quadratic* polynomials $f(x) = a_0 + a_1x + a_2x^2$ 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 $\mathrm{Ba(IO_3)_2}$ 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 $D(f) = \mathbb{Q}$ and if $m \neq 0$ then $R(f) = \mathbb{Q}$ 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.

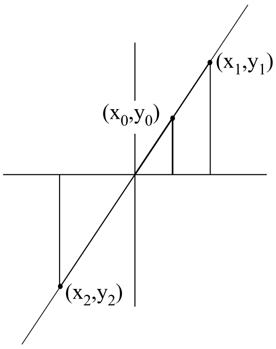


Fig. 10.1. Points satisfying $y = mx$ form similar triangles. In this figure, $m = 3/2$

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 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 $y_1 - y_0$ is the amount of "rise" corresponding to the "run" of $x_1 - x_0$. Hence the

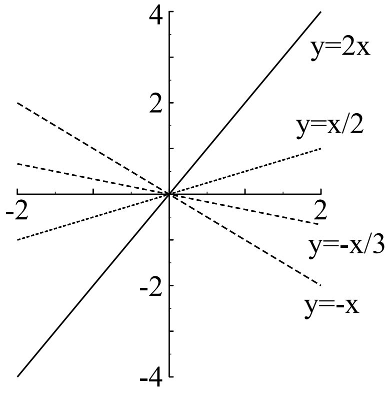


Fig. 10.2. Examples of lines

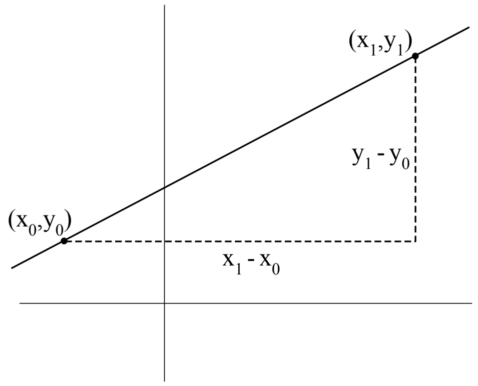


Fig. 10.3. The slope of any line is determined by the amount of rise over the amount of run between any two points on the line

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

$$(y - y_0) = m(x - x_0). \qquad (10.2)$$

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 line down. Evidently, we can find the slope-intercept form directly from knowing two points.

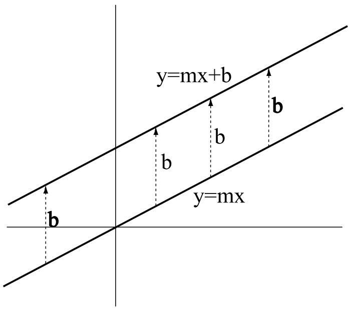


Fig. 10.4. The graph of $y = mx + b$ is found by translating the graph of $y = mx$ vertically by an amount b. In this case $b > 0$

Example 10.2. We find the slope-intercept form 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 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.

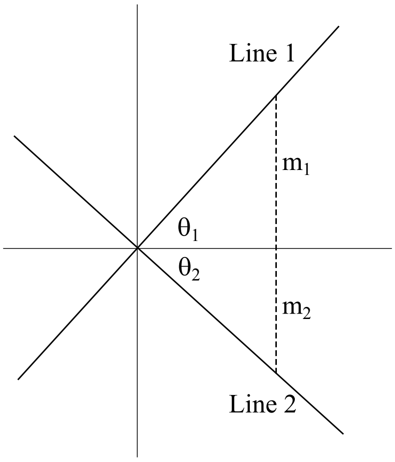


Fig. 10.5. Similar triangles defined by perpendicular lines with slope m_1 and m_2. The angles θ_1 and θ_2 add up to 90°

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 $m_1 m_2 < 0$.

Finally, assuming that $m_1 m_2 = -1$ 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 horizontal lines, will then be included in a natural way.

10.5 Quadratic Polynomials

The general quadratic polynomial has the form

$$f(x) = a_2 x^2 + a_1 x + a_0$$

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
-2	4
-1	1
$-.5$	.25

x	x^2
$-.25$	.125
$-.1$	.01
0	0

x	x^2
.1	.01
.5	.25
1	1

x	x^2
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 $x = 3$.

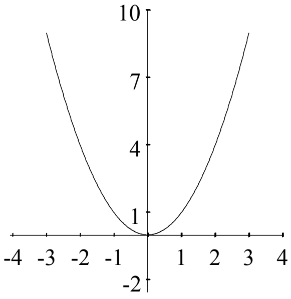


Fig. 10.6. Plot of $f(x) = x^2$. The function is decreasing for $x < 0$ and increasing for $x > 0$

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.

The second change we consider is translation. The two possibilities are to translate horizontally, or sideways, and vertically. We show examples of

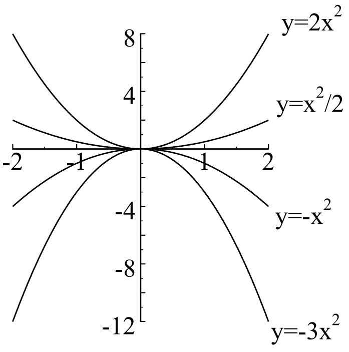


Fig. 10.7. Plots of $y = x^2$ scaled four different ways

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.

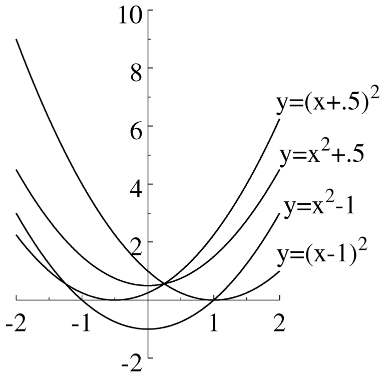


Fig. 10.8. Plots of $y = x^2$ translated four different ways

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).

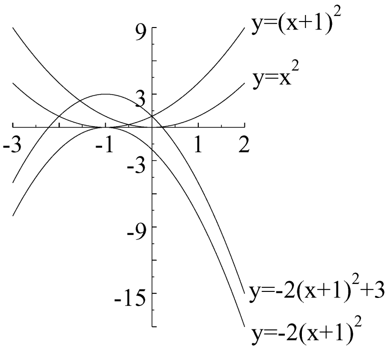


Fig. 10.9. Plotting $y = -2(x+1)^2 + 3$ in a systematic way

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 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

$$(x - x_0)^2 = x^2 - 2x_0x + x_0^2$$

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.

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.

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

$$a_i = \begin{cases} i + 2, & 0 \leq i \leq 15 \\ i, & 16 \leq i \leq 30 \end{cases}$$

Multiplication of a Polynomial by a Number

Given a polynomial

$$f(x)=\sum_{i=0}^{n}a_ix^i,$$

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:

$$(cf)(x)=\sum_{i=0}^{n}ca_ix^i.$$

We note that we can equivalently define $(cf)(x)$ by

$$(cf)(x)=cf(x)=c\times\ f(x).$$

Example 10.15.

$$2.3(1+6x-x^7)=2.3+13.8x-2.3x^7.$$

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.

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

$$f(x)=\sum_{m=1}^{n}c_mf_m(x)$$

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, $a_1, \ldots, a_n$. 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.

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

$$\begin{aligned}(fg)(x)=x^j g(x) &= a_0x^j + a_1x^j \times x + a_2x^j \times x^2 + \cdots + a_nx^j \times x^n\\ &= a_0x^j + a_1x^{1+j} + a_2x^{2+j} + \cdots + a_nx^{n+j}\\ &= \sum_{i=0}^{n} a_i x^{i+j},\end{aligned}$$

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

$$\begin{aligned}(fg)(x) = f(x)g(x) &= \left(\sum_{i=0}^{n} a_i x^i\right)\left(\sum_{i=0}^{m} b_i x^i\right)\\ &= \sum_{i=0}^{n}\left(a_i x^i \sum_{j=0}^{m} b_j x^j\right) = \sum_{i=0}^{n}\left(a_i \sum_{j=0}^{m} b_j x^{i+j}\right)\\ &= \sum_{i=0}^{n}\sum_{j=0}^{m} a_i b_j x^{i+j}.\end{aligned}$$

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:

$$\begin{aligned}(x + a)^2 = (x + a)(x + a) &= x^2 + 2ax + a^2\\ (x + a)(x - a) &= x^2 - a^2\\ (x + a)^3 &= x^3 + 3ax^2 + 3a^2x + a^3\end{aligned}$$

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, 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 $x = 3$.

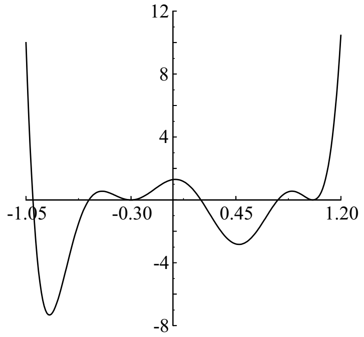


Fig. 10.10. A plot of $y = 1.296 + 1.296x - 35.496x^2 - 57.384x^3 + 177.457x^4$ $+203.889x^5 - 368.554x^6 - 211.266x^7 + 313.197x^8 + 70.965x^9 - 97.9x^{10} - 7.5x^{11}$ $+10x^{12}$

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

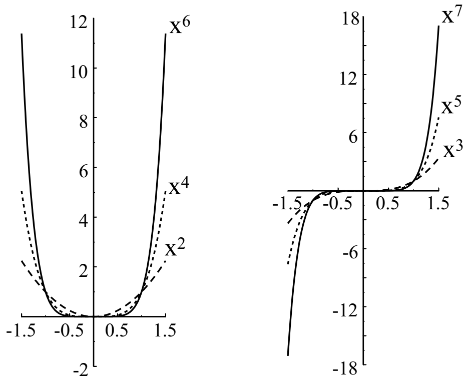


Fig. 10.11. Plots of some monomials

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.

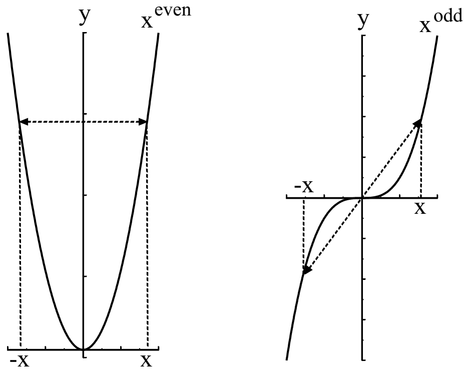


Fig. 10.12. The symmetries of the monomial functions of even and odd degree

We can use the ideas of scaling and translation to graph functions of the form $y = a(x - x_0)^m + d$.

Example 10.20. We plot $y = -.5(x-1)^3-6$ in Fig. 10.13 by systematically using translations and scaling. Luckily, however, there is no procedure like completing the square for monomials of higher degree.

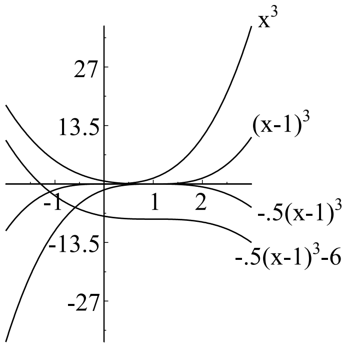


Fig. 10.13. The procedure for plotting $y = -.5(x-1)^3 - 6$

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.

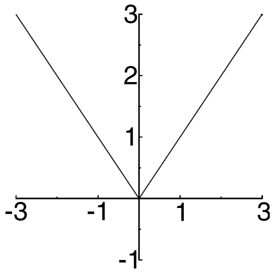


Fig. 10.14. Plot of $y = |x|$

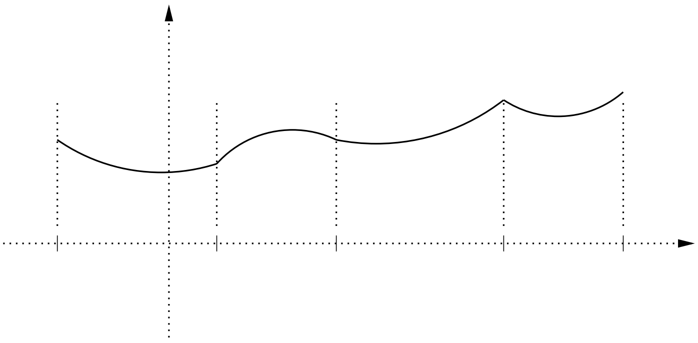


Fig. 10.15. Plot of a piecewise (quadratic) polynomial function

Chapter 10 Problems

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 $-2 \leq x \leq 2$: (a) $6x^2$, (b) $-\frac{1}{4}x^2$, (c) $\frac{4}{3}x^2$.

10.12. Plot the following quadratic functions for $-3 \leq x \leq 3$: (a) $(x - 2)^2$, (b) $(x + 1.5)^2$, (c) $(x + .5)^2$.

10.13. Plot the following quadratic functions for $-2 \leq x \leq 2$: (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_2f_1, (e) f_1f_3, (f) f_2f_3, (g) $f_1f_3-f_2$, (h) $(f_1+f_2)f_3$, (i) $f_1f_2f_3$.

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_1f_2 where $f_1(x)=\sum_{i=0}^{8} i^2x^i$ 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 $-3 \le x \le 3$: (a) x^3, (b) x^4, (c) x^5.

10.24. Plot the following polynomials for $-3 \le x \le 3$:

(a) $\frac{1}{3}(x+2)^3-2$ (b) $2(x-1)^4-13$ (c) $(x+1)^5-1$.

10.25. Plot the following piecewise polynomials for $-2 \le x \le 2$

(a) $f(x)=\begin{cases}1, & -2\le x\le -1,\\ x^2, & -1<x<1,\\ x, & 1\le x\le 2.\end{cases}$ (b) $f(x)=\begin{cases}-1-x, & -2\le x\le -1,\\ 1+x, & -1<x\le 0,\\ 1-x, & 0<x\le 1,\\ -1+x, & 1<x\le 2.\end{cases}$

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 $f_1 : D_{f_1} \to \mathbb{Q}, \dots, f_n : D_{f_n} \to \mathbb{Q}$, and numbers $c_1, \dots, c_n$, we define the *linear combination* of $f_1, \dots, f_n$ with coefficients $c_1, \dots, c_n$, 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 $\left\{\frac{1}{x}, \cdots, \frac{1}{x^n}\right\}$ 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

$$f(x) = (x^2 - 3)^3 \left(x^6 - \frac{1}{x} - 3 \right)$$

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 conclude that

$$\frac{x^3 + 4x^2 - 2x + 3}{x - 1} = x^2 + 5x + 3.$$

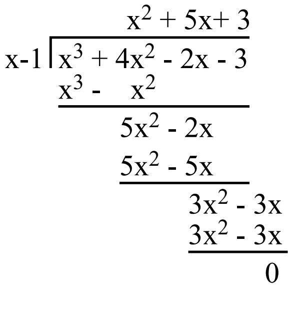


Fig. 11.1. An example of polynomial division with no remainder

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, the

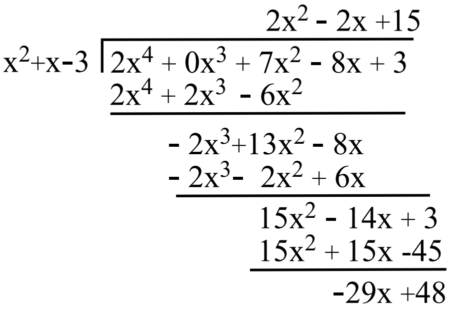


Fig. 11.2. An example of polynomial division with a remainder

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.

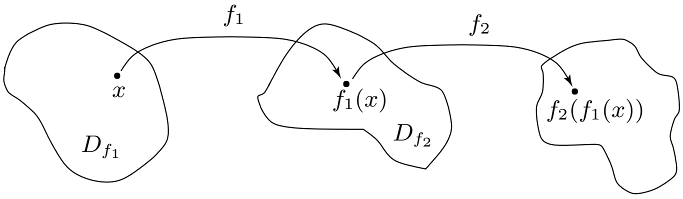


Fig. 11.3. Illustration of the composition $f_2 \circ f_1$

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.*

Fig. 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.

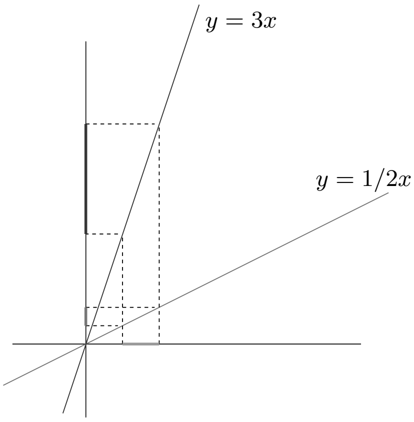


Fig. 12.2. These two linear functions which change a different amount for a given change in input

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

$$|x_2 + x_1| \le |x_2| + |x_1| \le 2 + 2 = 4,$$

by the triangle inequality. We conclude that

$$|f(x_2) - f(x_1)| \le 4|x_2 - x_1|$$

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

$$|x_2 + x_1| \leq |x_2| + |x_1| \leq 4 + 4 = 8$$

so

$$|f(x_2) - f(x_1)| \leq 8|x_2 - x_1|$$

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.

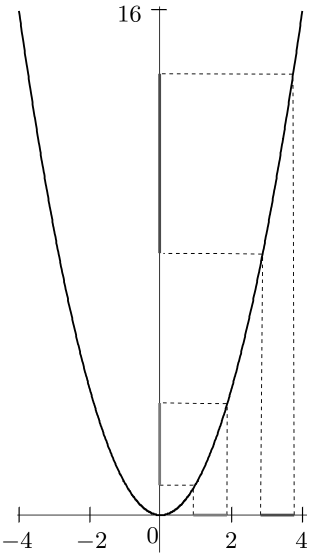


Fig. 12.3. The change in $f(x) = x^2$ for equal changes in x near $x = 2$ and $x = 4$

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 $|x| \leq a$ 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.

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

$$|f(x_2) - f(x_1)| = |x_2 + x_1||x_2 - x_1| \leq L_f|x_2 - x_1|$$

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 $L_f = m a^{m-1}$. 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

$$\left| \sum_{i=0}^{m-1} x_2^{m-1-i} x_1^i \right| \le \sum_{i=0}^{m-1} |x_2|^{m-1-i} |x_1|^i.$$

Now because x_1 and x_2 are in $[-a, a]$, $|x_1| \le a$ and $|x_2| \le a$. So

$$\left| \sum_{i=0}^{m-1} x_2^{m-1-i} x_1^i \right| \le \sum_{i=0}^{m-1} a^{m-1-i} a^i = \sum_{i=0}^{m-1} a^{m-1} = ma^{m-1}.$$

and

$$|f(x_2) - f(x_1)| \le ma^{m-1} |x_2 - x_1|.$$

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

$$\begin{aligned}|(f_1+f_2)(y)-(f_1+f_2)(x)| &= |(f_1(y)-f_1(x))+(f_2(y)-f_2(x))|\\ &\le |f_1(y)-f_1(x)|+|f_2(y)-f_2(x)|\\ &\le L_1|y-x|+L_2|y-x|\\ &= (L_1+L_2)|y-x|\end{aligned}$$

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 $c_1f_1 + c_2f_2$ is Lipschitz continuous with constant $|c_1|L_1 + |c_2|L_2$. Next given $i \le n$, 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

$$c_1f_1 + \cdots + c_if_i = \bigl(c_1f_1 + \cdots + c_{i-1}f_{i-1}\bigr) + c_nf_n.$$

But the assumption on $\bigl(c_1f_1 + \cdots + c_{i-1}f_{i-1}\bigr)$ means that we have written $c_1f_1 + \cdots + c_if_i$ as the sum of two Lipschitz continuous functions, namely $\bigl(c_1f_1 + \cdots + c_{i-1}f_{i-1}\bigr)$ 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

$$\begin{aligned}|f(x_2) - f(x_1)| &= \left|\left(x_2^4 - 3x_2^2\right) - \left(x_1^4 - 3x_1^2\right)\right| \\ &= \left|\left(x_2^4 - x_1^4\right) - \left(3x_2^2 - 3x_1^2\right)\right| \\ &\le \left|x_2^4 - x_1^4\right| + 3\left|x_2^2 - x_1^2\right|.\end{aligned}$$

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.

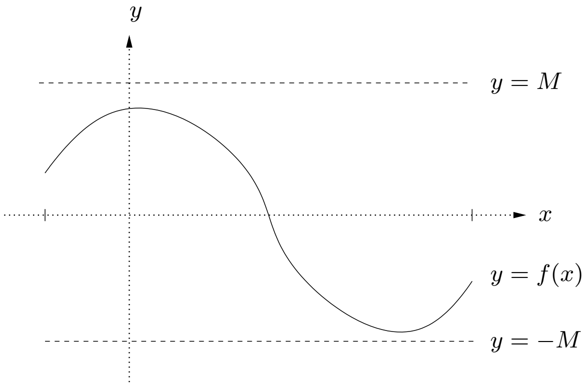


Fig. 12.4. A bounded function on I

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 $[-2, 2]$.

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

$$\frac{1}{x_2^2} - \frac{1}{x_1^2} = \frac{x_1^2}{x_1^2x_2^2} - \frac{x_2^2}{x_1^2x_2^2} = \frac{x_1^2 - x_2^2}{x_1^2x_2^2} = \frac{(x_1+x_2)(x_1-x_2)}{x_1^2x_2^2}.$$

This means that

$$|f(x_2) - f(x_1)| = \left|\frac{x_1+x_2}{x_1^2x_2^2}\right| |x_2 - x_1|.$$

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 functions 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

$$|f_2(f_1(y)) - f_2(f_1(x))| \leq L_2|f_1(y) - f_1(x)| \leq L_1L_2|y - x|.$$

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_1L_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

$$|f(x_1, y_1) - f(x_2, y_2)| \leq L_f(|x_1 - x_2| + |y_1 - y_2|)$$

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 $f(x_1, x_2) = x_1 x_2$ is Lipschitz continuous with Lipschitz constant $L_f = 2$, since for $x_1, x_2 \in [0, 1]$

$$|x_1x_2 - y_1y_2| = |x_1x_2 - y_1x_2 + y_1x_2 - y_1y_2|$$
$$\leq |x_1 - y_1|x_2 + y_1|x_2 - y_2| \leq 2(|x_1 - y_1| + |x_2 - y_2|).$$

12.11 Functions of Several Rational Variables

The concept of a function also extends to several variables, i.e. we consider functions $f(x_1, \ldots, x_d)$ 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 $f(x_1, \ldots, x_d)$ 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 $y_1, \cdots, y_d \in \mathbb{Q}$,

$$|f(x_1, \ldots, x_d) - f(y_1, \ldots, y_d)| \leq L_f(|x_1 - y_1| + \cdots + |x_d - y_d|).$$

Example 12.22. The function $f : \mathbb{R}^d \to \mathbb{Q}$ defined by $f(x_1, \ldots, x_d) = x_1 + x_2 + \cdots x_d$ 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 $1.11\ldots$ 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, $1.11\ldots 11_n$, and so on, as a *sequence* of numbers a_n, with $n = 1, 2, 3, \ldots$, where $a_1 = 1.1$, $a_2 = 1.11, \ldots, a_n = 1.11\ldots 11_n, \ldots$, 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 $1, 2, 3, \ldots$. 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 never ending list a_1, a_2, a_3, ..., of elements a_n, with the index n going through the natural numbers $n = 1, 2, 3, \ldots$. The symbol ∞, called "infinity", indicates that the list continues for ever in the same sense that the natural numbers $1, 2, 3, \ldots$, 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 $a_n = 1.11\ldots 11_n$, 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 never ending 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 $1.1, \ldots, 1.11..11_n, \ldots$, 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 $a_n = 1.11\ldots 11_n$, 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

$$1.111\ldots = \frac{10}{9},$$

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.

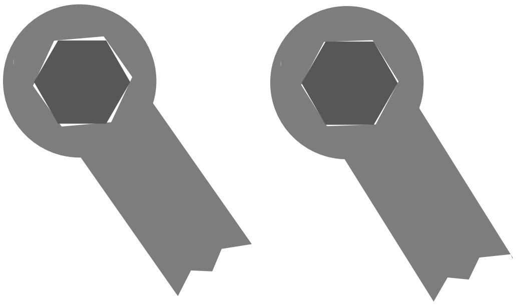


Fig. 13.1. Two socket wrenches with different tolerances

A *ideal expert mechanic* would have the whole sequence $\{0.66\cdots 67_n\}_{n=1}^{\infty}$ at his/her disposal with the error of wrench number n being estimated by

$$\left|0.66\cdots 67_n - \frac{2}{3}\right| \leq 10^{-n}.$$

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 $0.66\cdots 67_{20}$, 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 $0.6\cdots 67_n$ to increase (rapidly) with n, since the precision in the construction process has to improve.

As a general point, computing the numbers $0.66\cdots 67_n$ 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 $0.66\cdots 67_n$ 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.

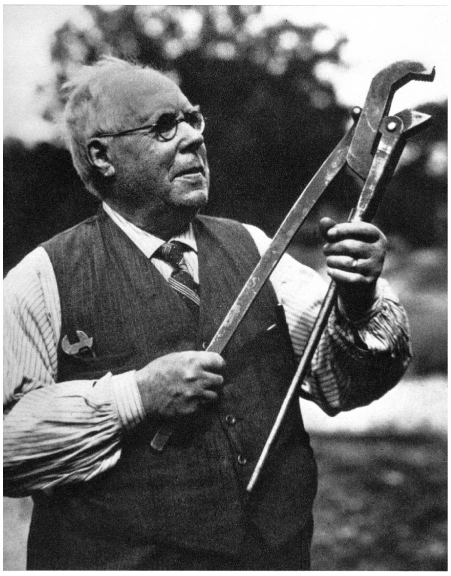


Fig. 13.2. The Swedish inventor J.P. Johansson with two adjustable wrenches of different design

13.4 The Power of Language: From Infinitely Many to One

The decimal expansion $0.6666\ldots$ of $\frac{2}{3}$ contains infinitely many decimals. The sequence $\{0.66\cdots 667_n\}_{n=1}^{\infty}$ 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 $0.6666\ldots$ is a "super-number" because it has infinitely many decimals, and then forget the "super" and say that $0.6666\ldots$, is a number. In fact, this makes complete sense since we identify $0.6666\ldots$ 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. 13.3 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

$$|a_n - A| \le \epsilon \quad \text{for all} \quad n \ge N$$

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

$$\left|\frac{10}{9} - 1.11\cdots 1_n\right| \le 10^{-n},$$

Fig. 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"

and thus

$$\left|\frac{10}{9} - 1.11\cdots 1_n\right| \leq \epsilon$$

if $10^{-n} \leq \epsilon$. We can phrase this as

$$\left|\frac{10}{9} - 1.11\cdots 1_n\right| \leq \epsilon$$

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.

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

$$|r|^{N+1} \leq \epsilon\,|1-r| \qquad (13.6)$$

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,

$$|a-b| \leq |a-c| + |c-b| \qquad (13.7)$$

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 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

$$\lim_{n\to\infty} f(a_n) = f\left(\lim_{n\to\infty} a_n\right).$$

The proof is also easy. By the Lipschitz continuity of $f : \mathbb{Q} \to \mathbb{Q}$, we have

$$\left| f(a_m) - f\left(\lim_{n\to\infty} a_n\right)\right| \leq L \left| a_m - \lim_{n\to\infty} a_n \right|.$$

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,*

$$\lim_{n\to\infty} f(a_n) = f\left(\lim_{n\to\infty} a_n\right). \qquad (13.8)$$

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

$$\lim_{n\to\infty} P_n = \lim_{n\to\infty} \frac{1}{\dfrac{1}{2^n} Q_0 + \dfrac{1}{K}\left(1 - \dfrac{1}{2^n}\right)}.$$

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 $\left\{\frac{1}{2^n}\right\}$. 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

$$\lim_{n\to\infty} (a_n)^2 = A^2.$$

In the next chapter, we will be interested in computing $\lim_{n\to\infty}(a_n)^2$ 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

$$f\left(\lim_{n\to\infty} a_n, \lim_{n\to\infty} b_n\right))^{\text{TS}^{\text{b}}} = \lim_{n\to\infty} f(a_n, b_n). \qquad (13.9)$$

The proof is immediate:

$$\left|f\left(\lim_{n\to\infty} a_n, \lim_{n\to\infty} b_n\right))^{\text{TS}^{\text{b}}} - f(a_m, b_m)\right| \leq L\left(\left|\lim_{n\to\infty} a_n - a_m\right| + \left|\lim_{n\to\infty} b_n - b_m\right|\right) \qquad (13.10)$$

TS^b Please check this closing parenthesis.

Editor's or typesetter's annotations (will be removed before the final TEX run)

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 $\{2 + 3n^{-4} + (-1)^n n^{-1}\}_{n=1}^\infty$.

$$\begin{aligned}\lim_{n\to\infty} &\left(2 + 3n^{-4} + (-1)^n n^{-1}\right)\\ &= \lim_{n\to\infty} 2 + 3 \lim_{n\to\infty} n^{-4} + \lim_{n\to\infty} (-1)^n n^{-1}\\ &= 2 + 3 \times 0 + 0 = 2.\end{aligned}$$

To do this example, we use (13.11) and the fact that

$$\lim_{n\to\infty} n^{-p} = \left(\lim_{n\to\infty} n^{-1}\right)^p = 0^p = 0$$

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

$$\begin{aligned} \lim_{n\to\infty} P_n &= \frac{1}{\displaystyle\lim_{n\to\infty} \frac{1}{2^n} Q_0 + \lim_{n\to\infty} \frac{1}{K}\left(1 - \frac{1}{2^n}\right)} \\ &= \frac{1}{0 + \frac{1}{K}(1-0)} = K. \end{aligned}$$

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:

$$\begin{aligned} \lim_{n\to\infty} 4\,\frac{1+n^{-3}}{3+n^{-2}} &= 4\,\frac{\lim_{n\to\infty}(1+n^{-3})}{\lim_{n\to\infty}(3+n^{-2})} \\ &= 4\,\frac{1+\lim_{n\to\infty} n^{-3}}{3+\lim_{n\to\infty} n^{-2}} \\ &= 4\,\frac{1+0}{3+0} = \frac{4}{3}. \end{aligned}$$

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 $0, 1, 2, \dots 9$. 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

$$\pm\, p_m 2^m + p_{m-1} 2^{m-1} + \ldots + p_2 2^2 + p_1 2^1 + p_0 2^0 + q_1 2^{-1} + q_2 2^{-2} \\ + \ldots + q_{n-1} 2^{-(n-1)} + q_n 2^{-n},$$

which we again may write in short hand

$$\pm p_m p_{m-1} \ldots p_1 p_0 . q_1 q_2 \ldots q_n = p_m p_{m-1} \ldots p_1 p_0 + 0.q_1 q_2 \ldots q_n$$

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 With 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. 21.34.

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.

Fig. 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)

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:

$$\text{(a)} \lim_{n\to\infty} \frac{8}{3n+1} = 0 \quad \text{(b)} \lim_{n\to\infty} \frac{4n+3}{7n-1} = \frac{4}{7} \quad \text{(c)} \lim_{n\to\infty} \frac{n^2}{n^2+1} = 1.$$

13.5. Show that $\lim_{n\to\infty} r^n = 0$ for any r with $|r| \leq 1/2$.

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 $|r| \geq 2$.

13.9. Find the values of

(a) $1 - .5 + .25 - .125 + \cdots$

(b) $3 + \frac{3}{4} + \frac{3}{16} + \cdots$

(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.

(a) $\left\{\dfrac{4^{n/2}}{4+(-1)^n}\right\}_{n=1}^{\infty}$ (b) $\left\{\dfrac{2^n}{4+(-1)^n}\right\}_{n=1}^{\infty}$

(c) $\left\{\dfrac{2^{\text{car}}}{4+(-1)^{\text{car}}}\right\}_{\text{car}=1}^{\infty}$ (d) $\left\{\dfrac{2^{n-1}}{4+(-1)^{n-1}}\right\}_{n=2}^{\infty}$

(e) $\left\{\dfrac{2^{n+2}}{4+(-1)^{n+2}}\right\}_{n=0}^{\infty}$ (f) $\left\{8\dfrac{2^n}{4+(-1)^{n+3}}\right\}_{n=-2}^{\infty}.$

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) $a_n = 1 + \frac{7}{n}$ (b) $a_n = 4n^2 - 6n$

(c) $a_n = \frac{(-1)^n}{n^2}$ (d) $a_n = \frac{2n^2 + 9n + 3}{6n^2 + 2}$

(e) $a_n = \frac{(-1)^n n^2}{7n^2 + 1}$ (f) $a_n = \left(\frac{2}{3}\right)^n + 2$

(g) $a_n = \frac{(n-1)^2 - (n+1)^2}{n}$ (h) $a_n = \frac{1 - 5n^8}{4 + 51n^3 + 8n^8}$

(i) $a_n = \frac{2n^3 + n + 1}{6n^2 - 5}$ (j) $a_n = \frac{\left(\frac{7}{8}\right)^n - 1}{\left(\frac{7}{8}\right)^n + 1}$.

13.17. Compute the following limits

(a) $\lim_{n\to\infty} \left(\frac{n+3}{2n+8}\right)^{37}$ (b) $\lim_{n\to\infty} \left(\frac{31}{n^2} + \frac{2}{n} + 7\right)^4$

(c) $\lim_{n\to\infty} \frac{1}{\left(2 + \frac{1}{n}\right)^8}$ (d) $\lim_{n\to\infty} \left(\left(\left(\left(1 + \frac{2}{n}\right)^2\right)^3\right)^4\right)^5$.

13.18. Rewrite the following sequences as a function applied to another sequence three different ways:

(a) $\left\{\left(\frac{n^2+2}{n^2+1}\right)^3\right\}_{n=1}^{\infty}$. (b) $\left\{(n^2)^4 + (n^2)^2 + 1\right\}_{n=1}^{\infty}$

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 Square Root 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

$$\begin{aligned} x^2 = 1.&999999999999999999999999999999999999999999999\\ &999999999999999999999999999999999999999999999\\ &999999999999999999999999999999999999999999999\\ &999999999999999999999999999999999999999999999\\ &999999999999999999999999999999999999999999999\\ &999999999999999999999999999999999999999999999\\ &999999999999999999999999999999999999999999999\\ &999999999999999999999999999999999999999999999\\ &999999999999999999999999999999999999999999999\\ &999999999986381037002790393547544921481567520\\ &719364336722392248627179189098787015809960232\\ &640597261312640760405691299950309295747831888\\ &596950070887405605833650165227157380944559332\\ &069004581726422217393596953324251515876023360\\ &427299488914180359897103820495618481233332162\\ &516016097283137123064499497943653479698629776\\ &683334066577024031851330600242723212517527304\\ &354776748660808998780793579777475964587708250\\ &317006887058548601 0 \end{aligned}$$

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 0*s* and 1*s*, 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

Fig. 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,

$$x_0 \le x_1 \le x_2 \le \cdots \quad \text{and} \quad X_0 \ge X_1 \ge X_2 \ge \cdots$$
$$x_i < X_j \quad \text{for all } i, j = 0, 1, 2, \ldots$$

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,

$$0 \leq X_i - x_i \leq 2^{-i} \quad \text{for } i = 0, 1, 2, \cdots, \tag{14.2}$$

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 $x_i \leq x_j < X_j \leq X_i$ if $j > i$ so (14.2) implies

$$|x_i - x_j| \leq |x_i - X_i| \leq 2^{-i} \quad \text{if } j \geq i.$$

that is

$$|x_i - x_j| \leq 2^{-i} \quad \text{if } j \geq i. \tag{14.3}$$

We illustrate in Fig. 14.2. In particular, this means that when $2^{-i} \leq 10^{-N-1}$, the first N decimals of x_j are the same as the first N decimals in x_i for any $j \geq i$.

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 $2^{-j} \leq 10^{-N-1}$. 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

$$\bar{x} = \lim_{i \to \infty} x_i.$$

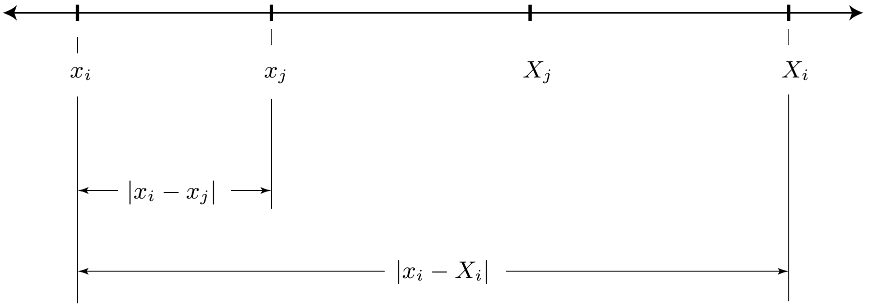


Fig. 14.2. $|x_i - x_j| \leq |X_i - x_i|$

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 algorithm 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*$^{\copyright}$ 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

Fig. 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. Solve the equations $x^2 = 3$ and $x^3 = 2$ using the Bisection algorithm. Also, make the algorithm find the negative root of $x^2 = 3$.

14.8. Show that if $a < 0$ and $b > 0$ then $b - a < c$ implies $|b| < c$ and $|a| < c$.

14.9. (a) Write down an algorithm for Deca-section. (b) Program the algorithm in (a) and then compute 16 steps using $x_0 = 0$ and $X_0 = 2$.

14.10. (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 $|x_i - \bar{x}|$.

14.11. Compute the cost of the tridiagonal algorithm from Problem 14.10 and compare to the costs of the Bisection and Deca-section methods.

14.12. Use the Bisection code from Problem 14.4 to compute $\sqrt{3}$ (recall Problem 3.5). Hint: $1 < \sqrt{3} < 2$.

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 never ending list of decimals $q_1, q_2, \ldots$, where each one of the p_i and q_j are one of the 10 digits $0, 1, \ldots, 9$. We met the decimal expansion

$1.4142135623\ldots$ 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 $\sqrt{2} = 1.412\ldots$, with $1.412\ldots$, 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 Chapter Do Mathematicians Quarrel? 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 $i = 1, 2, \ldots$, which defines the decimal expansion of $x + \bar{x}$, see Fig. 15.1. We may notice that occasionally adding two digits affects the digits to the left, as in $0.9999 + 0.0001 = 1.000$.

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*
⋮	⋮

Fig. 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

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 $i = 1, 2, \ldots$

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,

$$|x - x_i| \leq 10^{-i} \quad \text{for } i = 1, 2, \ldots \tag{15.4}$$

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

$$|x_i - \bar{x}_i| \leq 10^{-i} \quad \text{for } i = 1, 2, \ldots \tag{15.5}$$

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 Arithmetic 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

$$|f(x_i) - f(x_j)| \leq L|x_i - x_j|,$$

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 $\mathbb{R}$.

Similarly, we can generalize and extend a Lipschitz continuous function $f : I \to \mathbb{Q}$, where $I = \{x \in \mathbb{Q} : a \leq x \leq b\}$ is an interval of rational numbers, to a function $f : J \to \mathbb{R}$, where $J = \{x \in \mathbb{R} : a \leq x \leq b\}$ 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

$$|f(x) - f(y)| = \left|\lim_{i\to\infty} (f(x_i) - f(y_i))\right| \leq L \lim_{i\to\infty} |x_i - y_i| = L|x - y|.$$

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 see that the two functions agree at every point except for the “missing” point $x = 1$.

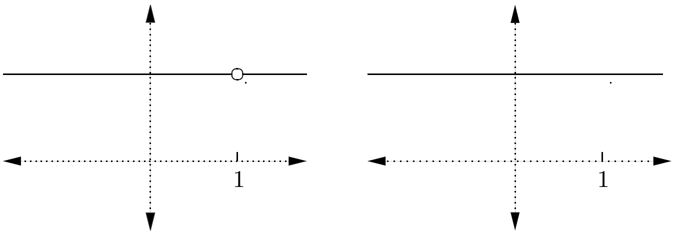


Fig. 15.2. Plots of $(x-1)/(x-1)$ on the *left* and 1 on the *right*

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 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 $\{x : a \le x \le b\}$ 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 $(a, b] = \{x : a < x \le b\}$. See Fig. 15.3.

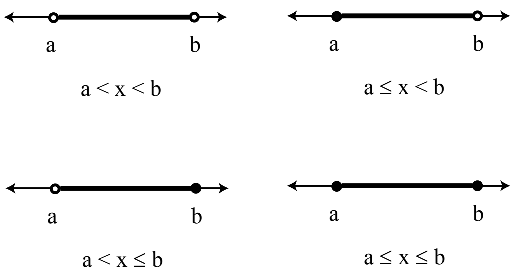


Fig. 15.3. Intervals corresponding to the real numbers between two real numbers a and b. Note the use of a solid and closed circles in the four cases

We also have "infinite" intervals such as $(-\infty, a) = \{x : x < a\}$ and $[b, \infty) = \{x : b \le x\}$. We illustrate these in Fig. 15.4. With this notation, 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 $\mathbb{R}$.

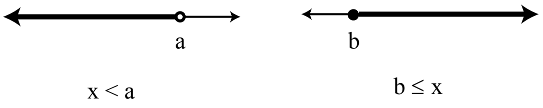


Fig. 15.4. Infinite intervals $(-\infty, a)$ and $[b, \infty)$

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 $i = 1, 2, \ldots$, 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	.1313707002030545241 5
9	1.41421356	.1313708442030479314744064
10	1.414213562	.1313708490030479221535281312
$\vdots$	$\vdots$	$\vdots$

Fig. 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.

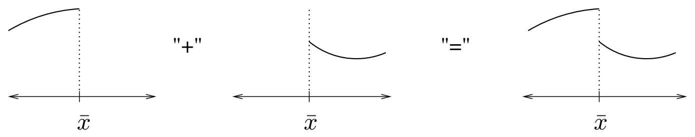


Fig. 15.6. A "jump function" viewed as two functions

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.

15.17 Continuity Versus 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 $f : I \to \mathbb{R}$:

$$f\left(\lim_{i\to\infty} x_i\right) = \lim_{i\to\infty} f(x_i), \tag{15.15}$$

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, $|x_i| \leq |x| + 1$.

15.3. Find i as small as possible such that $|xy - x_iy_i| \leq 10^{-6}$ if $x \approx 100$ and $y \approx 1$. Find i and j as small as possible such that $|xy - x_iy_j| \leq 10^{-6}$

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 $\left\{\dfrac{i}{i+1}\right\}$. 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) $|\sqrt{2}x-3|\leq 7$ 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.

$$\text{(a)}\ \left\{\frac{1}{(i+1)^2}\right\}\qquad \text{(b)}\ \left\{4-\frac{1}{2^i}\right\}\qquad \text{(c)}\ \left\{\frac{i}{3i+1}\right\}$$

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

$$\begin{aligned}
x_1 &= .373373337\cdots\\
x_2 &= .337733377333377\cdots\\
x_3 &= .333777333377733333777\cdots\\
x_4 &= .333377773333377773333337777\cdots\\
&\ \ \vdots
\end{aligned}$$

(a) Show that the sequence is a Cauchy sequence and (b) find $\lim\limits_{i\to\infty} x_i$. 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 $f(x) = y$.*

Fig. 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

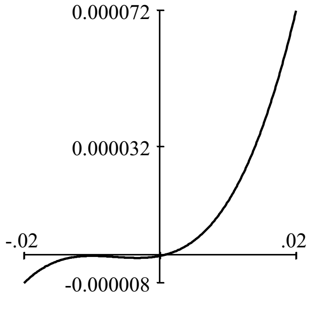


Fig. 16.2. A plot of the function $S\,(.02+2S)^2 - 1.57\times 10^{-9}$

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 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 algorithm can be used to compute its roots. We plot $f(x)$ in Fig. 16.3. It appears that $f(x) = 0$ might have one root near 0 and another root near -10.

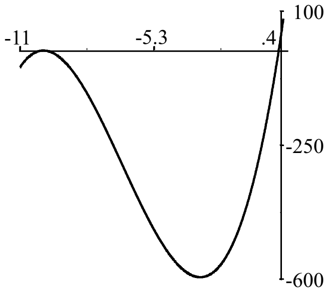


Fig. 16.3. A plot of the function $f(x) = x(20 + 2x)^2 - 1.57$

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

Fig. 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, Poincaré 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 and 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, dissertation at Kings College, Cambridge 1935) 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, seeTS^{c} Fig. 17.2. Turing laid out plans for an electronic computer named Analytical Computing Engine ACE, with reference

TS^{c} Figure 1 is not referred to in the text.

Editor's or typesetter's annotations (will be removed before the final TEX run)

Fig. 17.1. Kurt Gödel (with Einstein 1950): “Every formal system is incomplete”

Fig. 17.2. Alan Turing: “I wonder if my machine will come to a halt?”

to Babbages’ Analytical Engine, at the same time as the ENIAC was designed in the US.

Gödel's and Turing's work signified 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 and Turing's findings on the incompleteness of axiomatic methods and un-computability, 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 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) 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

Fig. 17.3. Bertrand Russell: "I am protesting"

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.

Fig. 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.

WILDROOT
The PATRIOTICK BARBER of NEW YORK, or the CAPTAIN in the SUDS.
Plate III

Fig. 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 $1, 2, 3, \ldots$, 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 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.

Fig. 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"

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 convinced 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 Versus 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 $1, 2, 3, 4, \ldots$, 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.

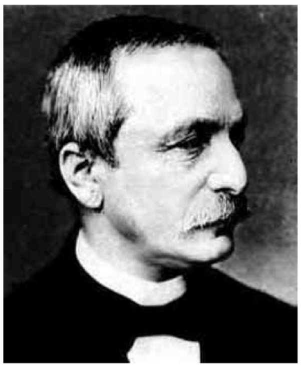


Fig. 17.7. Cantor (*left*): "I realize that in this undertaking I place myself in a certain opposition to views widely held concerning the mathematical infinite and to opinions frequently defended on the nature of numbers". Kronecker (*right*): "God created the integers, all else is the work of man"

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 enterprize 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 $n = 1, 2, \ldots$, and there is sTS$^{\text{d}}$ constant C such that $a_n \leq C$ for $n = 1, 2, \ldots$. 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.

TS$^{\text{d}}$ Please check “s”.

Editor’s or typesetter’s annotations (will be removed before the final TEX run)

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 $1/2, 3/4, 7/8, 15/16, 31/32, \ldots$, 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.

Turing's principle of a "universal computing machine" directly connects the work on the foundations of mathematics in the 1930s (with *Computable numbers* as a key article), with the development of the modern computer in the 1940s (with ACE as a key example), and thus very concretely illustrates the power of (constructive!) mathematics.

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)

My prize book at Sherbourne School (von Neumann's Mathematische Grundlagen der Quantenmechanik) is turning out very interesting, and not at all difficult reading, although the applied mathematicians seem to find it rather strong. (Turing, age 21)

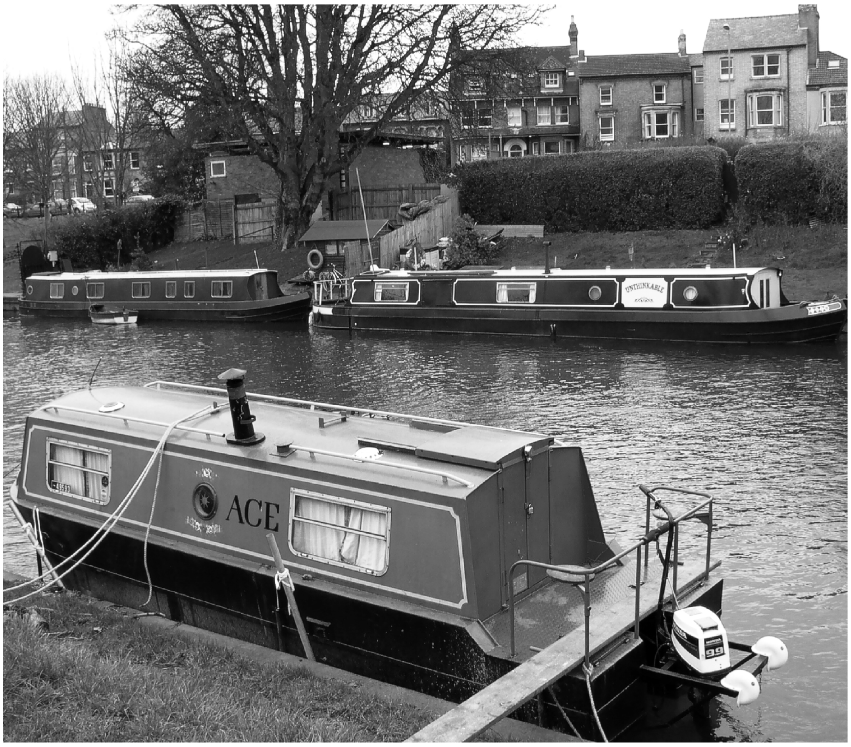


Fig. 17.8. View of the river Cam at Cambridge 2003 with ACE in the fore-ground (and "UNTHINKABLE" in the background to the right)

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

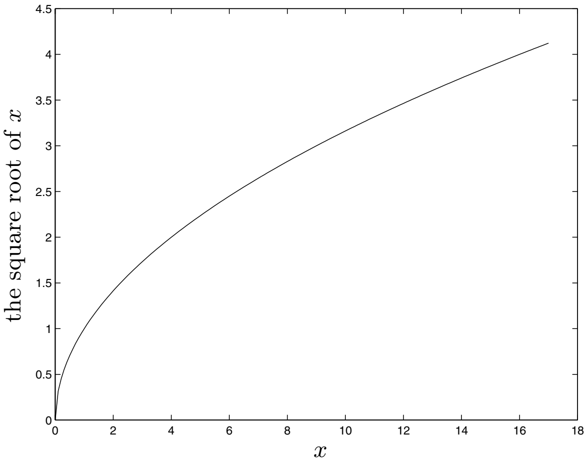


Fig. 18.1. The function $\sqrt{x}$ of x

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 for Rational $r = \frac{p}{q}$

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 (or *Hölder continuous* with exponent θ and constant L with a common terminology).

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!

18.7 Turbulent Flow is Hölder (Lipschitz) Continuous with Exponent $\frac{1}{3}$

In Chapter *Navier-Stokes, Quick and Easy* we give an argument indicating that turbulent flow is Hölder (Lipschitz) continuous with exponent $\frac{1}{3}$ so that a turbulent velocity $u(x)$ would satisfy

$$|u(x) - u(y)| \sim L|x - y|^{\frac{1}{3}}.$$

Such a turbulent velocity is a quite "wild" function which varies very quickly. Thus, Nature is not unfamiliar with Hölder (Lipschitz) continuity with exponent $\theta < 1$.

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?

19

Fixed Points and Contraction Mappings

> 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

$$\bar{x} = g(\bar{x}), \tag{19.1}$$

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, for $i = 1, 2, \ldots$, compute

$$x_i = g(x_{i-1}) \quad \text{for } i = 1, 2, 3, \ldots\ . \tag{19.2}$$

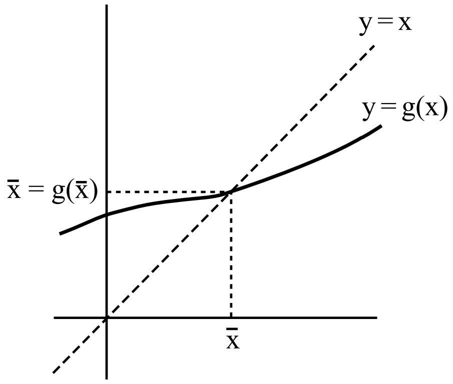


Fig. 19.1. Illustration of a fixed point problem $g(\bar{x}) = \bar{x}$

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 $x > y$,

$$\begin{aligned} g(x) - g(y) = x + \alpha f(x) - (y + \alpha f(y)) &= x - y + \alpha(f(x) - f(y)) \\ &= \left(1 + \alpha \frac{f(x) - f(y)}{x - y}\right) |x - y|, \end{aligned}$$

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.

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.

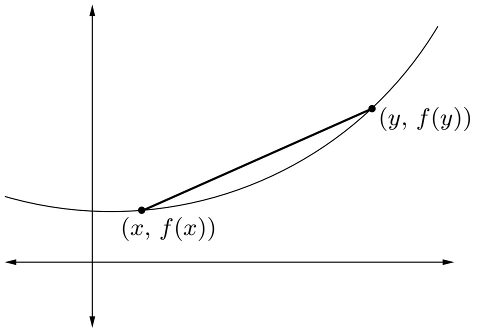


Fig. 19.2. Corda connecting the points $(x, f(x))$ and $(y, f(y))$ in $\mathbb{R}^2$

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

$$g(x) = \frac{1}{4} + \frac{1}{4}x \tag{19.4}$$

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.

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 $\bar{x} = 1/3$.

TS^e Figures 19.3 and 19.4 differ from printout, is this ok?

Editor's or typesetter's annotations (will be removed before the final TEX run)

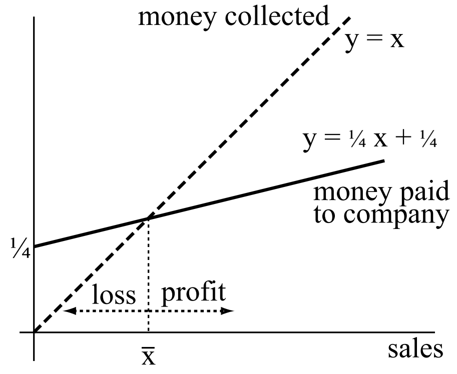


Fig. 19.3. Illustration of the problem of determining the break-even point for selling greeting cards door-to-door. Sales above the break-even point $\bar{x}$ give a profit to the salesman, but sales below this point mean a lossTS^{e}

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 $g(x_2) = \dots$ 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.

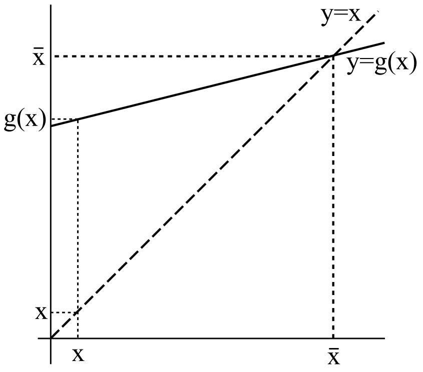


Fig. 19.4. The first step of Fixed Point Iteration in Card Sales model: $g(x)$ is closer to $\bar{x}$ than xTS$^{\text{e}}$

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 $\bar{x} = 1/3$,

$$|g(x_0) - \bar{x}| = \left|\frac{1}{4}x_0 + \frac{1}{4} - \frac{1}{3}\right| = \left|\frac{1}{4}\left(x_0 - \frac{1}{3}\right)\right| = \frac{1}{4}|x_0 - \bar{x}|.$$

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

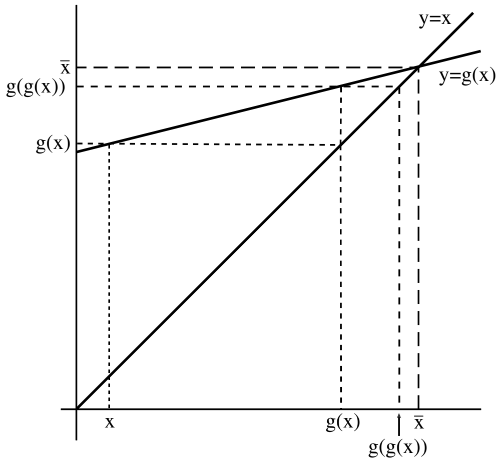


Fig. 19.5. Two steps of the contraction map algorithm applied to the fixed point problem in Model 19.4. The distance of $g(g(x))$ to $\bar{x}$ is 1/4 the distance from $g(x)$ to $\bar{x}$ and 1/16 the distance from x to $\bar{x}$

$$|x_i - \bar{x}| = \frac{1}{4}|x_{i-1} - \bar{x}|,$$

and thus for $i = 1, 2, \ldots,$

$$|x_i - \bar{x}| = 4^{-i}|x_0 - \bar{x}|.$$

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 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 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.

	Bisection Algorithm	Fixed Point Iteration
i	X_i	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	

Fig. 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

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

Fig. 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 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 $L < 1$.

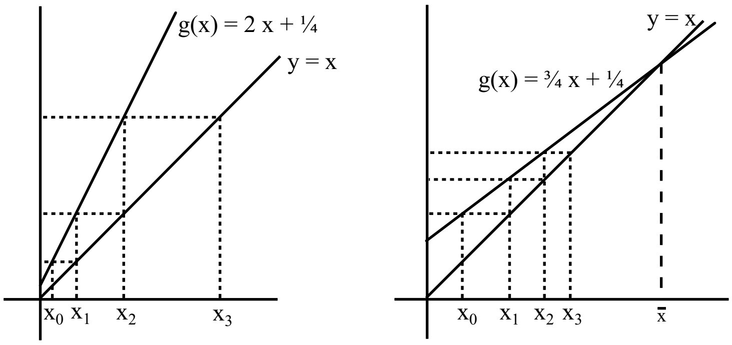


Fig. 19.8. On the *left*, we plot the first three fixed point iterates for $g(x) = 2x + \frac{1}{4}$. The iterates increase without bound as the iteration proceeds. On the *right*, we plot the first three fixed point iterates for $g(x) = \frac{3}{4}x + \frac{1}{4}$. The iteration converges to the fixed point in this case

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

$$|x_{k+1} - x_k| \leq L|x_k - x_{k-1}|. \tag{19.7}$$

Similarly,

$$|x_k - x_{k-1}| \leq L|x_{k-1} - x_{k-2}|,$$

and thus

$$|x_{k+1} - x_k| \leq L^2 |x_{k-1} - x_{k-2}|.$$

Repeating the argument, we find that

$$|x_{k+1} - x_k| \leq L^k |x_1 - x_0|. \tag{19.8}$$

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,

$$|x_i - x_j| \leq |x_i - x_{i+1}| + |x_{i+1} - x_{i+2}| + \cdots + |x_{j-1} - x_j| = \sum_{k=i}^{j-1} |x_k - x_{k+1}|.$$

We now use (19.8) on each term $|x_k - x_{k+1}|$ in the sum to get

$$|x_i - x_j| \leq \sum_{k=i}^{j-1} L^k \, |x_1 - x_0| = |x_1 - x_0| \sum_{k=i}^{j-1} L^k.$$

We compute

$$\sum_{k=i}^{j-1} L^k = L^i \left(1 + L + L^2 + \cdots + L^{j-i-1}\right) = L^i \, \frac{1 - L^{j-i}}{1 - L},$$

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

$$g\left(\bar{x}\right) = g\left(\lim_{i \to \infty} x_i\right) = \lim_{i \to \infty} g\left(x_i\right).$$

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 $\frac{1}{9}x + \frac{3}{4}$

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

Fig. 19.9. Results of the fixed point iterations for $\frac{1}{9}x + \frac{3}{4}$ and $\frac{1}{5}x + 2$

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 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 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.

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

Fig. 19.10. Results of the fixed point iterations for $\frac{1}{2}x$ and $\frac{1}{2}x^2$

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 accuracy. The convergence appears to be very quick.

i	x_i
0	1.00000000000000
1	1.50000000000000
2	1.41666666666667
3	1.41421568627451
4	1.41421356237469
5	1.41421356237310
6	1.41421356237310

Fig. 19.11. The fixed point iteration for (19.10)

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

$$\begin{aligned}\left|x_i-\sqrt{2}\right| &= \left|g\left(x_{i-1}\right)-g\left(\sqrt{2}\right)\right| \\ &= \left|\frac{x_{i-1}}{2}+\frac{1}{x_{i-1}}-\left(\frac{\sqrt{2}}{2}+\frac{1}{\sqrt{2}}\right)\right| \\ &= \left|\frac{x_{i-1}^2+2}{2x_{i-1}}-\sqrt{2}\right|.\end{aligned}$$

Now we find a common denominator for the fractions on the right and then use the fact that

$$\left(x_{i-1}-\sqrt{2}\right)^2 = x_{i-1}^2-2\sqrt{2}x_{i-1}+2$$

to get

$$\left|x_i-\sqrt{2}\right| = \frac{\left(x_{i-1}-\sqrt{2}\right)^2}{2x_{i-1}} \approx \frac{1}{2\sqrt{2}}\left(x_{i-1}-\sqrt{2}\right)^2. \tag{19.11}$$

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!

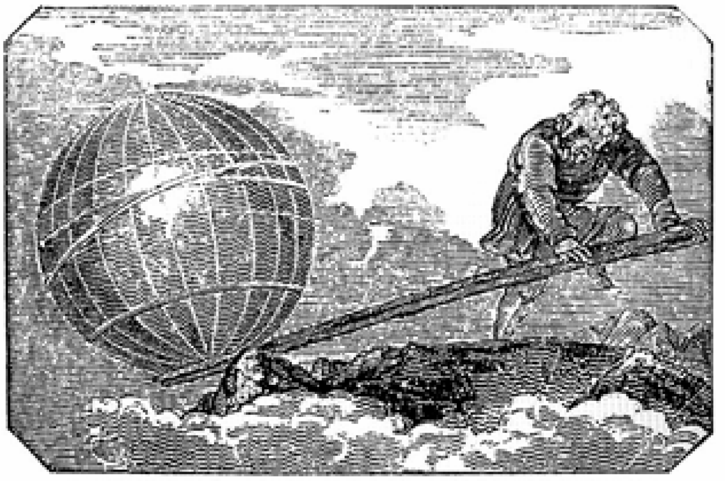


Fig. 19.12. Archimedes moving the Earth with a lever and a fixed point

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: $1.8\times10^{-5} = 18\times10^{-6}$ 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 19.2 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 (19.8).

19.18. Verify the details of Example 19.4.

19.19. (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.

19.20. Verify that (19.11) is true.

19.21. (a) Show the root problem $f(x) = x^2 + x - 6$ can be written as the fixed point problem $g(x) = x$ with $g(x) = \frac{6}{x+1}$. 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 $g(x) = \frac{x^2+6}{2x+1}$. 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.

19.22. 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.

19.23. 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 méthode pour bien conduire sa raison et chercher la vérité dans les sciences* from 1637. *The Method* contained as an appendix *La Géometrie* 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, might or might not become aware of the differential outflow of *animal*

Fig. 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"

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 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 $\mathbb{R}^2$

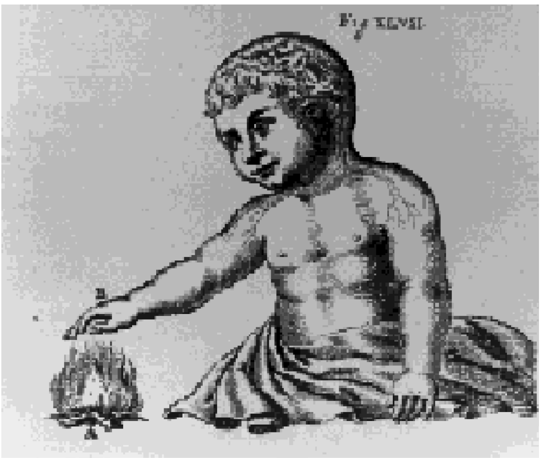


Fig. 20.2. Automatic reaction in response to external stimulation from Descartes *De homine* 1662

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.

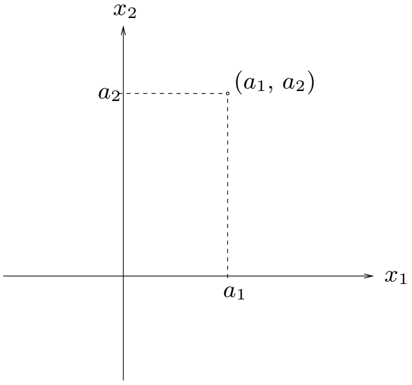


Fig. 20.3. Coordinate system for $\mathbb{R}^2$

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.

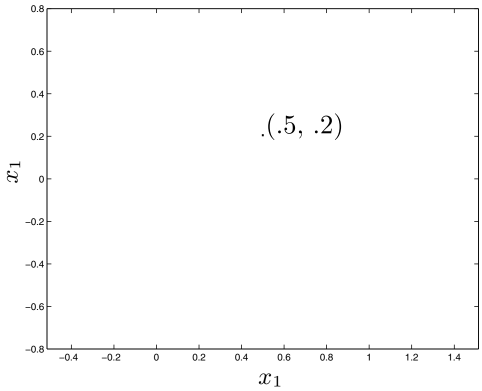


Fig. 20.4. Matlabs way of visualizing a coordinate system for a plane

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.

The wind arrow 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 $(8\sin(22.5^\circ), 8\cos(22.5^\circ)) \approx (3.06, 7.39)$.

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. 20.5. 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.

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

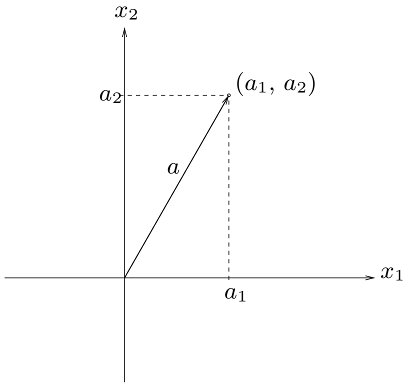


Fig. 20.5. A vector with tail at the origin and the head at the point $a = (a_1, a_2)$

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 $\underline{\mathrm{a}}$, 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 Vector Addition and 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. 20.6. 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. 20.6.

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. 20.6.

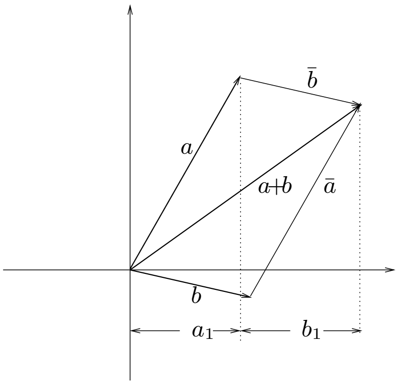


Fig. 20.6. Vector addition using the Parallelogram Law

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. 20.7.

$$(a_1, a_2) = (a_1, 0) + (0, a_2). \tag{20.2}$$

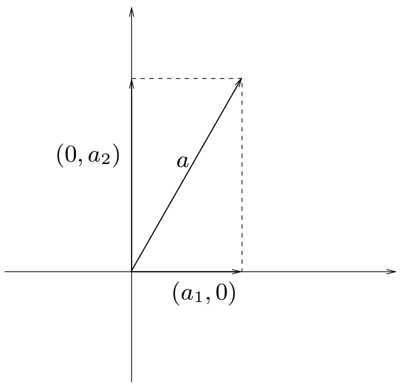


Fig. 20.7. A vector a represented as the sum of two vectors parallel with the coordinate axes

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.

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. 20.8.

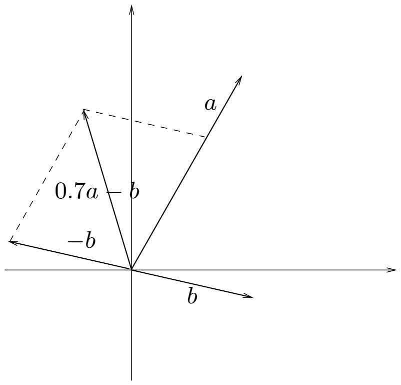


Fig. 20.8. The sum $0.7a - b$ of the multiples $0.7a$ and $(-1)b$ of a and b

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

$$|a| = \left(a_1^2 + a_2^2\right)^{1/2}. \tag{20.4}$$

By Pythagoras theorem and Fig. 20.9, 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.

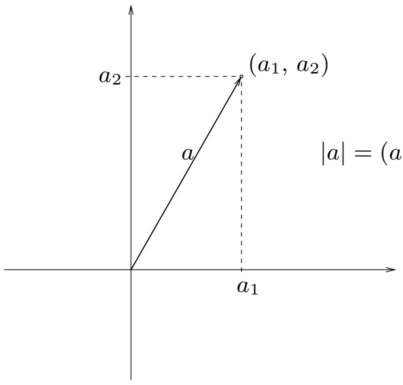


Fig. 20.9. The norm $|a|$ of a vector $a = (a_1, a_2)$ is $|a| = (a_1^2 + a_2^2)^{1/2}$

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. 20.10.

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. 20.10

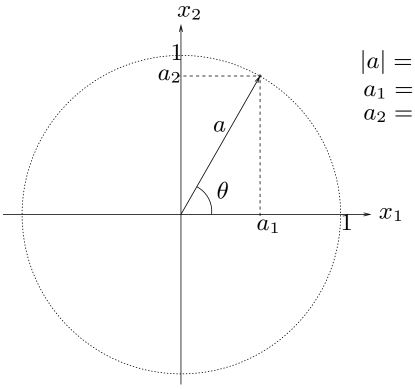


Fig. 20.10. Vectors of length one are given by $(\cos(\theta), \sin(\theta))$

Any vector $a = (a_1, a_2) \neq (0, 0)$ can be expressed as

$$a = |a|\hat{a} = r(\cos(\theta), \sin(\theta)), \tag{20.5}$$

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. 20.11. We call (20.5) the *polar representation* of a. We call θ the direction of a and r the length of a, see Fig. 20.11.

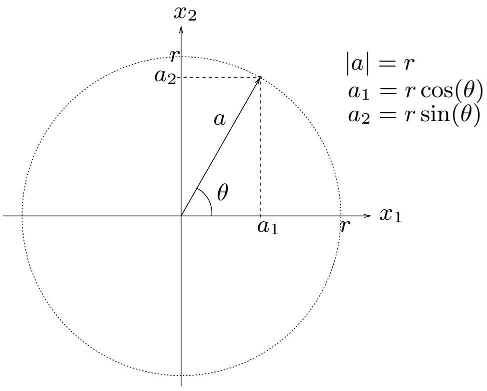


Fig. 20.11. Vectors of length r are given by $a = r(\cos(\theta), \sin(\theta)) = (r\cos(\theta), r\sin(\theta))$ where $r = |a|$

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

$$(1,1) = \sqrt{2}(\cos(45°), \sin(45°)) \text{ and } (-1,1) = \sqrt{2}(\cos(135°), \sin(135°)).$$

20.13 Standard Basis Vectors

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

$$a_1 e_1 + a_2 e_2 = a_1(1,0) + a_2(0,1) = (a_1, 0) + (0, a_2) = (a_1, a_2) = a.$$

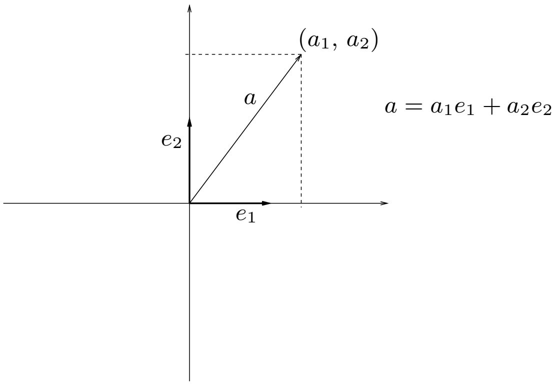


Fig. 20.12. The standard basis vectors e_1 and e_2 and a linear combination $a = (a_1, a_2) = a_1 e_1 + a_2 e_2$ of e_1 and e_2

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. 20.12.

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 $a. * b = (a_1 b_1, a_2 b_2)$.)

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:

$$\begin{aligned}a \cdot b &= (a_1e_1 + a_2e_2) \cdot (b_1e_1 + b_2e_2)\\ &= a_1b_1\, e_1 \cdot e_1 + a_1b_2\, e_1 \cdot e_2 + a_2b_1\, e_2 \cdot e_1 + a_2b_2\, e_2 \cdot e_2 = a_1b_1 + a_2b_2.\end{aligned}$$

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. 20.13. 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. 20.13. 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.

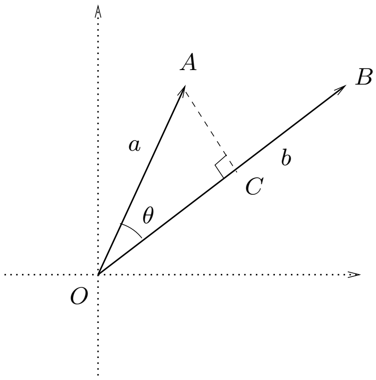


Fig. 20.13. $a \cdot b = |a|\,|b|\cos(\theta)$

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°,

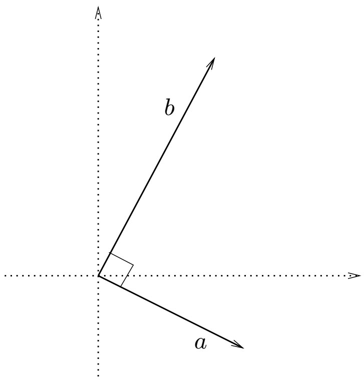


Fig. 20.14. Orthogonal vectors a and b

see Fig. 20.14. The basis vectors e_1 and e_2 are examples of geometrically orthogonal vectors, see Fig. 20.12.

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 $B = (b_1, b_2)$ 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.

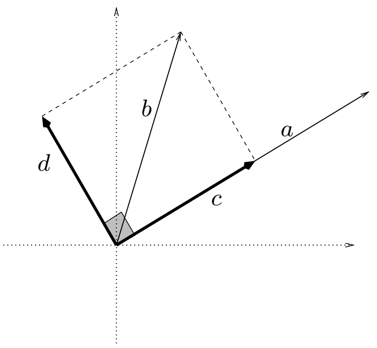


Fig. 20.15. Orthogonal decomposition of b

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. 20.15. 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:

$$P_a(b) = \lambda a \quad \text{for some } \lambda \in \mathbb{R},$$
$$(b - P_a(b)) \cdot a = 0.$$

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

$$P_a(b) = \frac{b \cdot a}{|a|^2}\, a. \tag{20.9}$$

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 Sect. 20.16, see also Fig. 20.15.

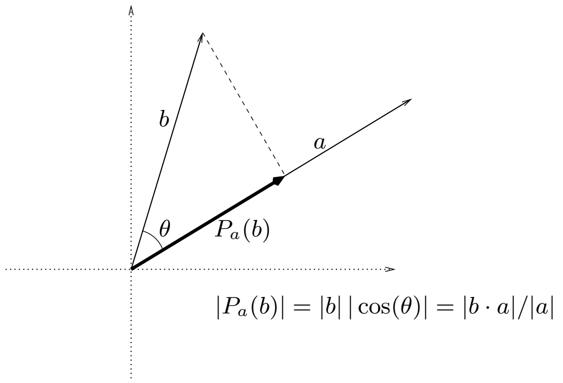


Fig. 20.16. The projection $P_a(b)$ of b onto a

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. 20.17.

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!

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

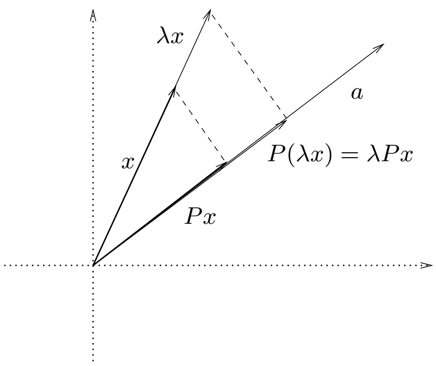


Fig. 20.17. $P(\lambda x) = \lambda Px$

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. 20.18. Similarly, the vector $(a_2, -a_1) = -(-a_2, a_1)$ is obtained by clockwise rotation of (a_1, a_2) by 90°.

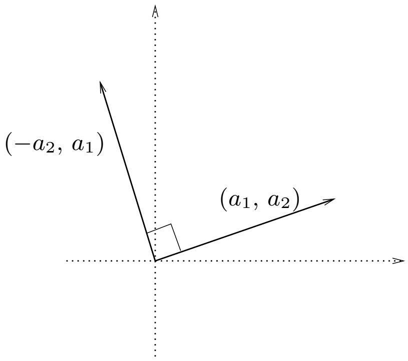


Fig. 20.18. Counter-clockwise rotation of $a = (a_1, a_2)$ by 90°

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

$$\begin{gathered} f(a+b) = (-(a_2+b_2), a_1+b_1) = (-a_2, a_1) + (-b_2, b_1) = f(a) + f(b), \\ f(\lambda a) = (-\lambda a_2, \lambda a_1) = \lambda(-a_2, a_1) = \lambda f(a). \end{gathered}$$

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

$$\begin{aligned} a^\perp &= f(a) = f(a_1e_1 + a_2e_2) = a_1f(e_1) + a_2f(e_2) \\ &= a_1(0,1) + a_2(-1,0) = (-a_2, a_1). \end{aligned}$$

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,

$$\begin{aligned} \cos(\alpha+\theta) &= \cos(\alpha)\cos(\theta) - \sin(\alpha)\sin(\theta), \\ \sin(\alpha+\theta) &= \sin(\alpha)\cos(\theta) + \cos(\alpha)\sin(\theta), \end{aligned}$$

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,

$$\text{(i)} \quad |R_\theta(a)| = |a|, \quad \text{and} \quad \text{(ii)} \quad R_\theta(a) \cdot a = \cos(\theta)|a|^2. \qquad (20.15)$$

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

$$\begin{aligned} a_1e_1 + a_2e_2 &= \hat{a}_1\hat{e}_1 + \hat{a}_2\hat{e}_2 \\ &= \hat{a}_1(\cos(\theta)e_1 + \sin(\theta)e_2) + \hat{a}_2(-\sin(\theta)e_1 + \cos(\theta)e_2) \\ &= (\hat{a}_1\cos(\theta) - \hat{a}_2\sin(\theta))e_1 + (\hat{a}_1\sin(\theta) + \hat{a}_2\cos(\theta))e_2, \end{aligned}$$

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

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

$$a \times b = a_1b_2 - a_2b_1. \tag{20.18}$$

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 "×" 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 $\mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$. 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,

$$\begin{aligned} a \times b &= (a_1e_1 + a_2e_2) \times (b_1e_1 + b_2e_2) = a_1b_2e_1 \times e_2 + a_2b_1e_2 \times e_1 \\ &= a_1b_2 - a_2b_1e_2. \end{aligned}$$

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 $\theta + 90°$:

$$\begin{aligned} |a \times b| = \left|a^\perp \cdot b\right| &= \left|a^\perp\right| |b| \left|\cos\left(\theta + \frac{\pi}{2}\right)\right| \\ &= |a|\,|b|\,|\sin(\theta)|, \end{aligned}$$

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. 20.19.

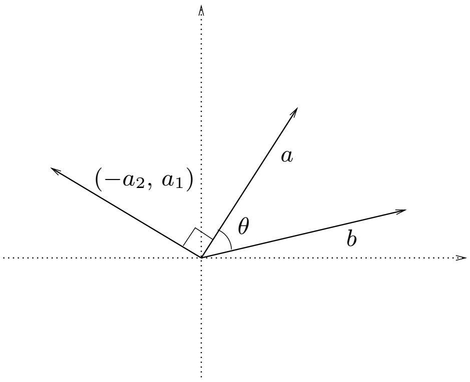


Fig. 20.19. Why $|a \times b| = |a|\,|b|\,|\sin(\theta)|$

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. 20.20.

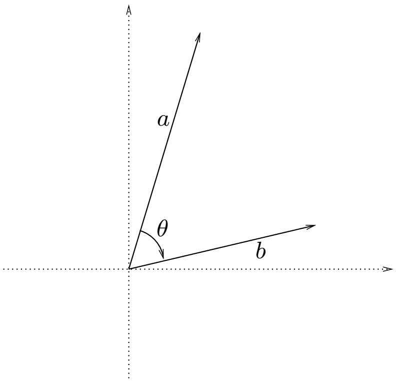


Fig. 20.20. $a \times b = |a|\,|b| \sin(\theta)$ is negative here because the angle θ is negative

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.

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. 20.21. 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. 20.21. Recalling (20.21), we conclude

Theorem 20.6

$$\mathit{Area}(OAB) = \frac{1}{2}|a|\,|b|\,|\sin(\theta)| = \frac{1}{2}\,|a \times b|.$$

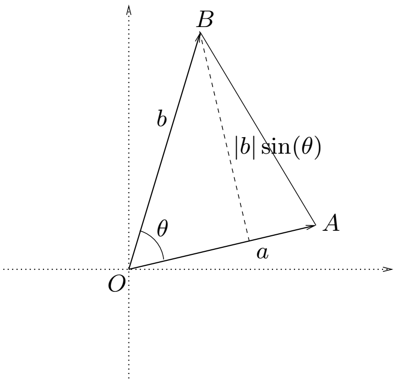


Fig. 20.21. The vectors a and b span a triangle with area $\frac{1}{2}|a \times b|$

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. 20.22.

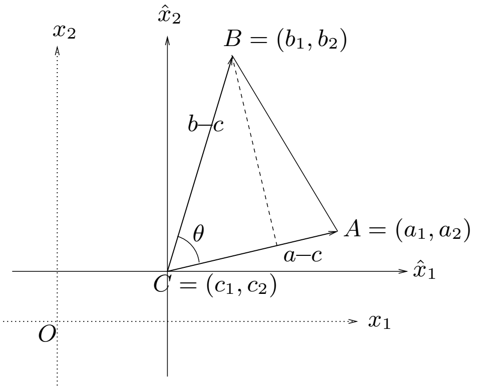


Fig. 20.22. Vectors $a - c$ and $b - c$ span triangle with area $\frac{1}{2}|(a - c) \times (b - c)|$

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 the result from the previous section, we find the area of the triangle CAB by the formula

$$\text{Area}(CAB) = \frac{1}{2} \left|(a - c) \times (b - c)\right|. \qquad (20.23)$$

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. 20.23, 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

$$V(a, b) = |a \times b|. \qquad (20.24)$$

This is a fundamental formula which has important generalizations to $\mathbb{R}^3$ and $\mathbb{R}^n$.

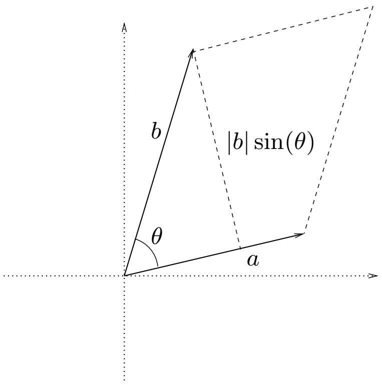


Fig. 20.23. The vectors a and b span a rectangle with area $|a \times b| = |a|\,|b|\,\sin(\theta)|$

20.27 Straight Lines

The points $x = (x_1, x_2)$ in the plane $\mathbb{R}^2$ satisfying a relation of the form

$$n_1x_1 + n_2x_2 = n \cdot x = 0, \qquad (20.25)$$

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. 20.24. 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. 20.24. We state this insight as a theorem because of its importance.

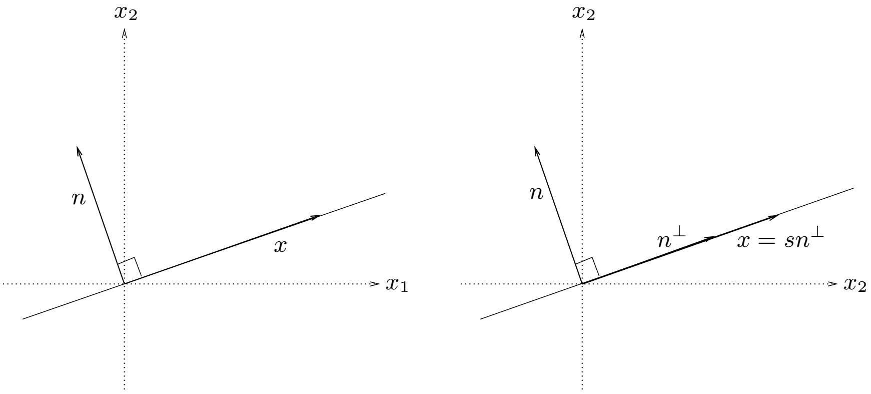


Fig. 20.24. Vectors $x = sa$ with b orthogonal to a given vector n generate a line through the origin with normal a

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 $s \in \mathbb{R}$.*

Similarly, the set of points (x_1, x_2) in $\mathbb{R}^2$ such that

$$n_1x_1 + n_2x_2 = d, \tag{20.26}$$

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. 20.25. 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 $n \cdot x = d$.

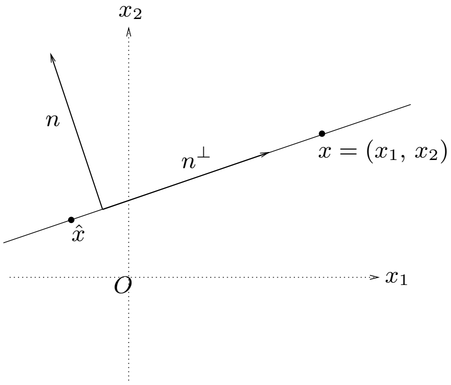


Fig. 20.25. The line through the point $\hat{x}$ with normal n generated by directional vector a

The line in Fig. 20.24 can also be represented as the set of points

$$x = \hat{x} + sn^\perp \quad s \in \mathbb{R},$$

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 $x = sn^\perp + \hat{x}$ 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 $x = \hat{x} + sn^\perp$ for $s \in \mathbb{R}$, where $\hat{x} \in \mathbb{R}^2$ is a point on the line.*

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

$$\begin{aligned} n \cdot Pb = d \quad &(Pb \text{ is a point on the line}), \\ b - Pb \text{ is parallel to the normal } n, \quad &(b - Pb = \lambda n, \quad \text{for some } \lambda \in \mathbb{R}). \end{aligned}$$

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

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 &= b_1, \\ a_{21}x_1 + a_{22}x_2 &= b_2, \end{aligned}$$

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

$$x_1a_1 + x_2a_2 = b. \tag{20.32}$$

Taking the vector product of this equation with a_2 and a_1 and using $a_2 \times a_2 = a_1 \times a_1 = 0$,

$$x_1a_1 \times a_2 = b \times a_2, \quad x_2a_2 \times a_1 = b \times a_1.$$

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 $x_1a_1 + x_2a_2 = b$, 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 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

$$x_1a_1 + x_2a_2 = 0$$

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 $x_1a_1 + x_2a_2 = 0$, 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 $a_1 \times a_2 \neq 0$.*

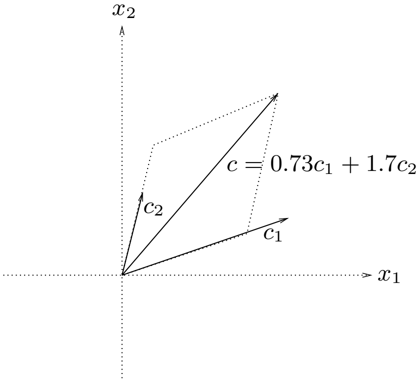


Fig. 20.26. Linear combination c of two linearly independent vectors c_1 and c_2

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 $x_1a_1 + x_2a_2 = b$, 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 $a_1x_1 + a_2x_2 = b$, 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 Mappings $f : \mathbb{R}^2 \to \mathbb{R}$

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:

$$f(x) = x_1c_1 + x_2c_2 = c_1x_1 + c_2x_2,$$

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 Mappings $f : \mathbb{R}^2 \to \mathbb{R}^2$

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. 20.27.

Example 20.16. The function $f(x_1, x_2) = (x_1 + 3x_2, 2x_1 - x_3)$ defines a linear mapping $\mathbb{R}^2 \to \mathbb{R}^2$.

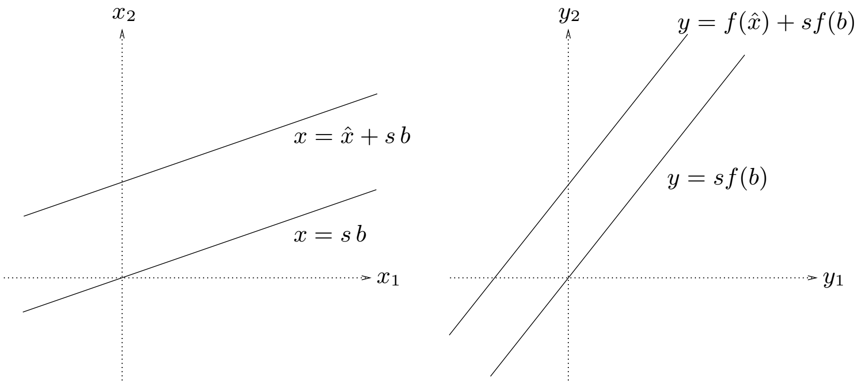


Fig. 20.27. A linear mapping $f : \mathbb{R}^2 \to \mathbb{R}^2$ maps (parallel) lines to (parallel) lines, and consequently parallelograms to parallelograms

20.35 Linear Mappings and Linear Systems 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 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:*

$$Ax = b, \qquad \text{or} \quad \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

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 $R_\theta : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ 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

$$\begin{aligned} AB = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \\ = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix} = C. \end{aligned}$$

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

$$B=\begin{pmatrix}\frac{a_1^2}{|a|^2}&\frac{a_1a_2}{|a|^2}\\\frac{a_1a_2}{|a|^2}&\frac{a_2^2}{|a|^2}\end{pmatrix}=\frac{1}{|a|^2}\begin{pmatrix}a_1^2&a_1a_2\\a_1a_2&a_2^2\end{pmatrix}.$$

We have

$$BB=\frac{1}{|a|^4}\begin{pmatrix}a_1^2&a_1a_2\\a_1a_2&a_2^2\end{pmatrix}\begin{pmatrix}a_1^2&a_1a_2\\a_1a_2&a_2^2\end{pmatrix}$$

$$=\frac{1}{|a|^4}\begin{pmatrix}a_1^2(a_1^2+a_2^2)&a_1a_2(a_1^2+a_2^2)\\a_1a_2(a_1^2+a_2^2)&a_2^2(a_1^2+a_2^2)\end{pmatrix}=\frac{1}{|a|^2}\begin{pmatrix}a_1^2&a_1a_2\\a_1a_2&a_2^2\end{pmatrix}=B,$$

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:

$$A=\begin{pmatrix}\cos(\alpha)&-\sin(\alpha)\\\sin(\alpha)&\cos(\alpha)\end{pmatrix}\quad\text{and}\quad B=\begin{pmatrix}\cos(\beta)&-\sin(\beta)\\\sin(\beta)&\cos(\beta)\end{pmatrix}.\qquad(20.44)$$

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:

$$\begin{gathered} (A+B)^\top = A^\top + B^\top, \quad (\lambda A)^\top = \lambda A^\top, \\ (AB)^\top = B^\top A^\top. \end{gathered}$$

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:

$$\begin{aligned} (\lambda A)^{-1} &= \lambda A^{-1} \\ (AB)^{-1} &= B^{-1}A^{-1}. \end{aligned}$$

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. 20.28, 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.

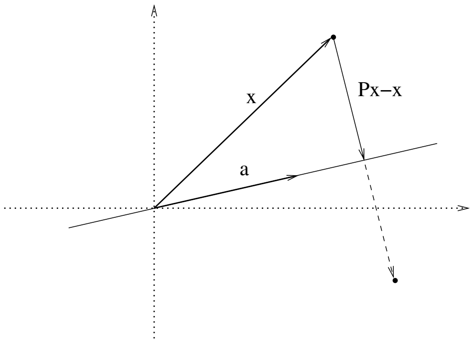


Fig. 20.28. The mapping $2P - I$ maps points to its mirror point relative to the given line

To see if $2P - I$ preserves scalar products, we assume that $y = (2P - I)x$ and $\hat{y} = (2P - I)\hat{x}$ and compute:

$$y \cdot \hat{y} = ((2P - I)x)^\top (2P - I)\hat{x} = x^\top (2P^\top - I)(2P - I)\hat{x} = \qquad (20.52)$$
$$x^\top (4P^\top P - 2P^\top I - 2PI + I)\hat{x} = x^\top (4P - 4P + I)\hat{x} = x \cdot \hat{x}, \qquad (20.53)$$

where we used the fact that $P = P^\top$ and $PP = P$, and we thus find an affirmative answer.

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

$$b = x_1 a_1 + x_2 a_2 = \hat{x}_1 \hat{a}_1 + \hat{x}_2 \hat{a}_2, \qquad (20.54)$$

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 $\{a_1, a_2\}$:

$$c_{11} a_1 + c_{21} a_2 = \hat{a}_1,$$
$$c_{12} a_1 + c_{22} a_2 = \hat{a}_2,$$

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,

$$(c_{11} \hat{x}_1 + c_{12} \hat{x}_2) a_1 + (c_{21} \hat{x}_1 + c_{22} \hat{x}_2) a_2 = b.$$

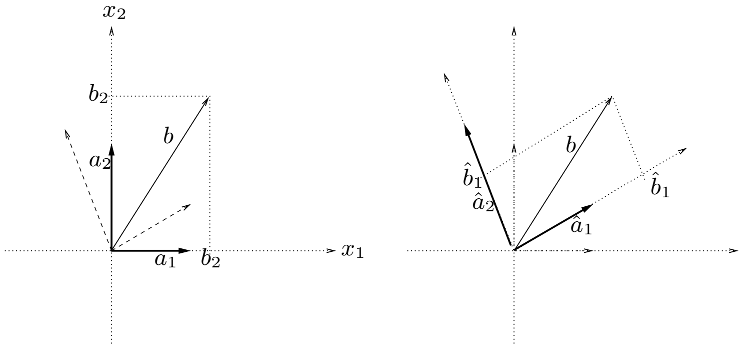


Fig. 20.29. A vector b may be expressed in terms of the basis $\{a_1, a_2\}$ or the basis $\{\hat{a}_1, \hat{a}_2\}$

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

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}.$$

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 philoso-

Fig. 20.30. 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"

phers, 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

20.1. Given the vectors a, b and c in $\mathbb{R}^2$ and the scalars $\lambda, \mu \in \mathbb{R}$, prove the following statements

$$\begin{aligned} a+b=b+a, \quad (a+b)+c=a+(b+c), \quad a+(-a)=0\\ a+0=a, \quad 3a=a+a+a, \quad \lambda(\mu a)=(\lambda\mu)a,\\ (\lambda+\mu)a=\lambda a+\mu a, \quad \lambda(a+b)=\lambda a+\lambda b, \quad |\lambda a|=|\lambda||a|. \end{aligned}$$

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. 20.17, 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)

> You can't help respecting anybody who can spell TUESDAY, even if he doesn't spell it right. (The House at Pooh Corner, Milne)

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

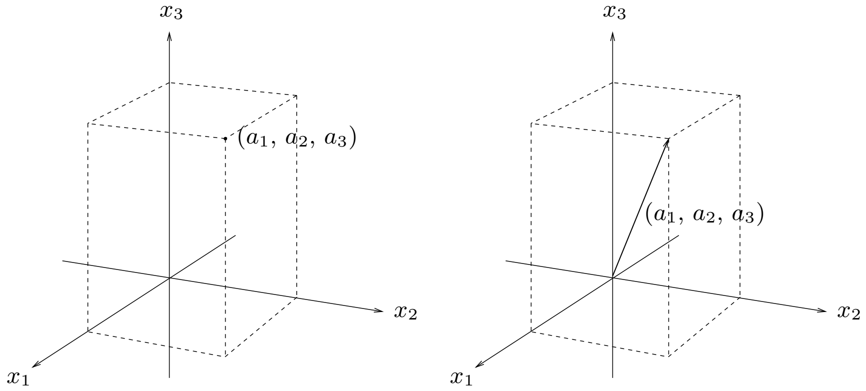


Fig. 21.1. Coordinate system for $\mathbb{R}^3$

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 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.

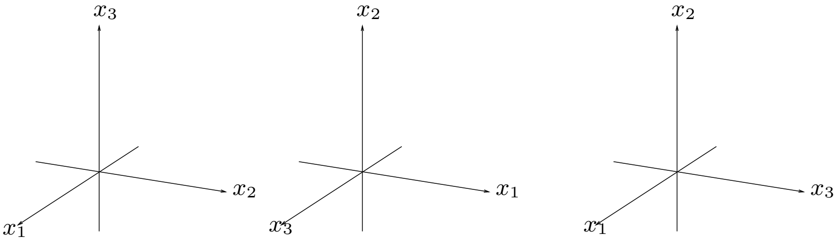


Fig. 21.2. Two "right" coordinate systems and one "left", where the vertical coordinate of the view point is assumed positive, that is, the horizontal plane is seen from above. What happens if the vertical coordinate of the view point is assumed negative?

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

$$|a| = (a \cdot a)^{\frac{1}{2}} = \left(\sum_{i=1}^{3} a_i^2\right)^{\frac{1}{2}}, \tag{21.2}$$

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

$$|a-b| = \left(\sum_{i=1}^{3}(a_i-b_i)^2\right)^{1/2}.$$

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 $P:\mathbb{R}^3\rightarrow\mathbb{R}^3$ 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

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 $a \cdot b = |a||b|\cos(\theta) = 0$.

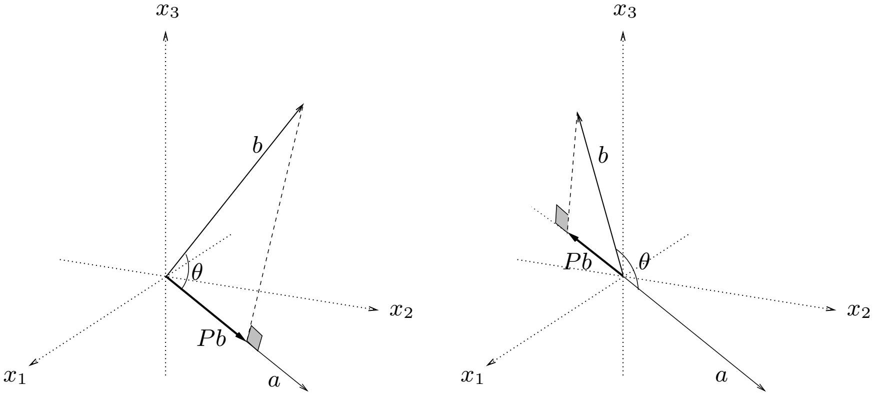


Fig. 21.3. Projection Pb of a vector b onto a vector a

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

$$a \times b = (a_2b_3 - a_3b_2, a_3b_1 - a_1b_3, a_1b_2 - a_2b_1)\,. \tag{21.7}$$

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) $a.*b = (a_1b_1, a_2b_2, a_3b_3)$. 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 $a_1b_2 - a_2b_1$ 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

$$a \cdot (a \times b) = a_1(a_2b_3 - a_3b_2) + a_2(a_3b_1 - a_1b_3) + a_3(a_1b_2 - a_2b_1) = 0, \tag{21.11}$$

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,

$$\begin{aligned} a \times b &= (a_1 e_1 + a_2 e_2 + a_3 e_3) \times (b_1 e_1 + b_2 e_2 + b_3 e_3) \\ &= a_1 b_2\, e_1 \times e_2 + a_2 b_1\, e_2 \times e_1 \\ &\qquad + a_1 b_3\, e_1 \times e_3 + a_3 b_1\, e_3 \times e_1 \\ &\qquad\qquad + a_2 b_3\, e_2 \times e_3 + a_3 b_2\, e_3 \times e_2 \\ &= (a_1 b_2 - a_2 b_1) e_3 + (a_3 b_1 - a_1 b_3) e_2 + (a_2 b_3 - a_3 b_2) e_1 \\ &= (a_2 b_3 - a_3 b_2, a_3 b_1 - a_1 b_3, a_1 b_2 - a_2 b_1), \end{aligned}$$

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 $a \times b = (0, 0, a_1 b_2 - a_2 b_1)$ 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 $a \times b = 0$.*

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.

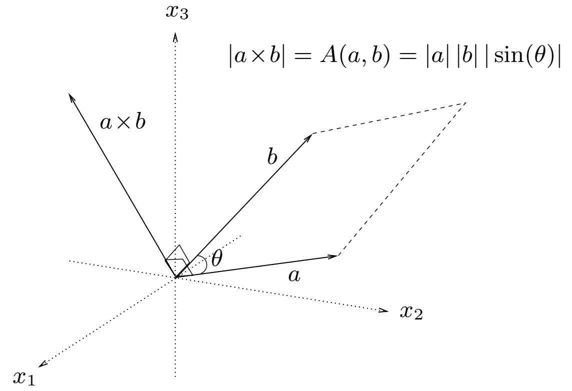


Fig. 21.4. Geometric interpretation of the vector product

21.8 Connection Between 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

$$a \times b = (0, 0, a_1 b_2 - a_2 b_1). \qquad (21.14)$$

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 a 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 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

$$A(a, b) = |a \times b|, \quad h = \frac{|c \cdot (a \times b)|}{|a \times b|},$$

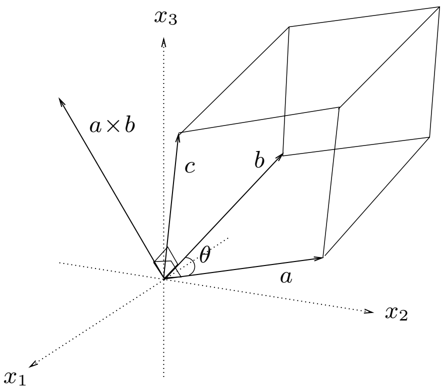


Fig. 21.5. Parallelepiped spanned by three vectors

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 × 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 × 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 A Formula for the Volume Spanned by Three Vectors

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 (21.18) 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

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.

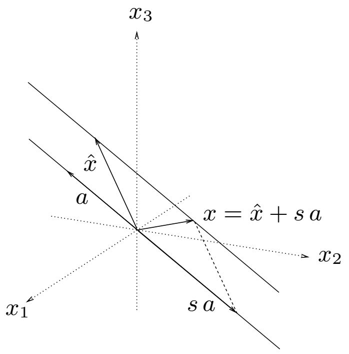


Fig. 21.6. Line in $\mathbb{R}^3$ of the form $x=\hat{x}+s\,a$

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.

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

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

$$n \cdot x = n \cdot \hat{x}. \tag{21.24}$$

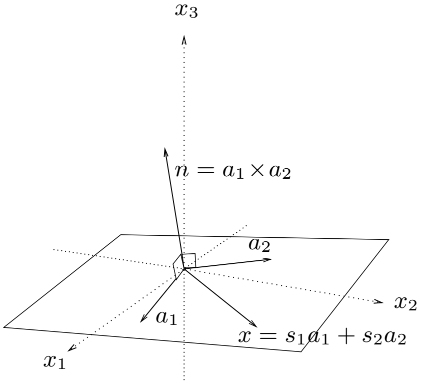


Fig. 21.7. Plane through the origin spanned by a_1 and a_2, and with normal $n = a_1 \times a_2$

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

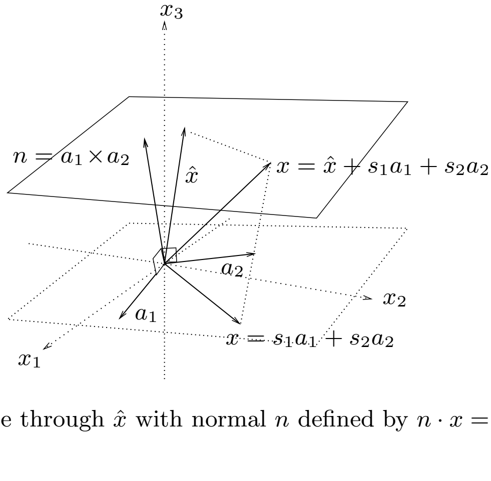


Fig. 21.8. Plane through $\hat{x}$ with normal n defined by $n \cdot x = d = n \cdot \hat{x}$

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. 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 Determine 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,

$$n \cdot Pb = d, \text{ that is } Pb \text{ is a point in the plane,}$$
$$b - Pb \text{ is parallel to the normal } n, \text{ that is } b - Pb = \lambda n \text{ for some } \lambda \in \mathbb{R}.$$

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

$$Pb = b - \frac{b \cdot n - d}{|n|^2} n. \qquad (21.28)$$

If $d = 0$ so the plane $n \cdot x = d = 0$ passes through the origin, then

$$Pb = b - \frac{b \cdot n}{|n|^2} n. \qquad (21.29)$$

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

$$\begin{aligned} x_1 a_1 \cdot a_1 + x_2 a_2 \cdot a_1 &= b \cdot a_1 - \hat{x} \cdot a_1, \\ x_1 a_1 \cdot a_2 + x_2 a_2 \cdot a_2 &= b \cdot a_2 - \hat{x} \cdot a_2 \end{aligned} \qquad (21.30)$$

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.

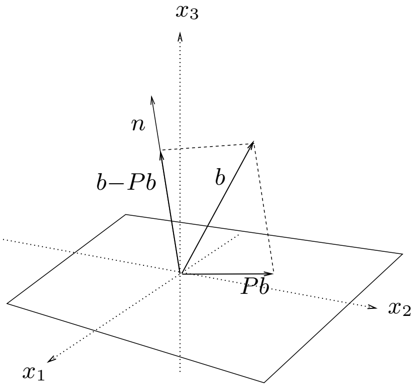


Fig. 21.9. Projection of a point/vector onto a plane

Example 21.10. The distance from the point $(2, 2, 3)$ to the plane $x_1+x_2+x_3 = 1$ is equal to $\frac{|(2,2,3)\cdot(1,1,1)-1|}{\sqrt{3}} = 2\sqrt{3}$.

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 $Rx - Px$ 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

$$Rx = Px + \cos(\theta)(x - Px) + \frac{\sin(\theta)}{|a|} a \times (x - Px). \qquad (21.31)$$

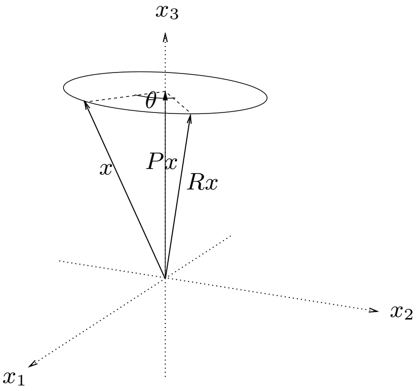


Fig. 21.10. Rotation around $a = (0, 0, 1)$ a given angle θ

21.20 Lines and Planes Through the Origin Are Subspaces

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 $\mathbb{R}^3$:

$$x_1a_1+x_2a_2+x_3a_3=b, \tag{21.33}$$

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

$$V(a_1,a_2,a_3)=|a_1\cdot a_2\times a_3|\neq 0, \tag{21.34}$$

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 $x_1a_1+x_2a_2+x_3a_3=b$ 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 $a_1\cdot a_2\times a_3=a_2\cdot a_3\times a_1=a_3\cdot a_1\times a_2$):

$$\begin{aligned}
x_1&=\frac{b\cdot a_2\times a_3}{a_1\cdot a_2\times a_3},\\
x_2&=\frac{b\cdot a_3\times a_1}{a_2\cdot a_3\times a_1}=\frac{a_1\cdot b\times a_3}{a_1\cdot a_2\times a_3},\\
x_3&=\frac{b\cdot a_1\times a_2}{a_3\cdot a_1\times a_2}=\frac{a_1\cdot a_2\times b}{a_1\cdot a_2\times a_3},
\end{aligned} \tag{21.35}$$

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

$$\begin{array}{rcl} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &=& b_1, \\ (a_{22}a_{11} - a_{21}a_{12})x_2 + (a_{23}a_{11} - a_{21}a_{13})x_3 &=& a_{11}b_2 - a_{21}b_1, \\ (a_{32}a_{11} - a_{31}a_{12})x_2 + (a_{33}a_{11} - a_{31}a_{13})x_3 &=& a_{11}b_3 - a_{31}b_1, \end{array} \tag{21.36}$$

where the unknown x_1 has been *eliminiated* in the second and third equations. This system has the form

$$\begin{array}{rrl} a_{11}x_1 + & a_{12}x_2 + a_{13}x_3 = & b_1, \\ & \hat{a}_{22}x_2 + \hat{a}_{23}x_3 = & \hat{b}_2, \\ & \hat{a}_{32}x_2 + \hat{a}_{33}x_3 = & \hat{b}_3, \end{array} \tag{21.37}$$

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

$$\begin{array}{rrl} a_{11}x_1 + & a_{12}x_2 + a_{13}x_3 = & b_1, \\ & \hat{a}_{22}x_2 + \hat{a}_{23}x_3 = & \hat{b}_2, \\ & \tilde{a}_{33}x_3 = & \tilde{b}_3, \end{array} \tag{21.38}$$

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

$$\begin{array}{rrl} x_1 & +2x_2 + 3x_3 = & 6, \\ 2x_1 & +3x_2 + 4x_3 = & 9, \\ 3x_1 & +4x_2 + 6x_3 = & 13. \end{array}$$

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

$$\begin{array}{rrl} x_1 & +2x_2 + 3x_3 = & 6, \\ & -x_2 - 2x_3 = & -3, \\ & -2x_2 - 3x_3 = & -5. \end{array}$$

Subtracting now the second equation multiplied by 2 from the third equation, we get

$$\begin{aligned} x_1 + 2x_2 + \quad 3x_3 &= \quad 6, \\ -x_2 - \quad 2x_3 &= \ -3, \\ x_3 &= \quad 1, \end{aligned}$$

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 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:

$$x_1a_1 + x_2a_2 + x_3a_3 = b, \tag{21.40}$$

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 (21.18) 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 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

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 $\{a_1, a_2, a_3\}$,

$$b = x_1a_1 + x_2a_2 + x_3a_3, \qquad (21.43)$$

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 $b = x_1a_1 + x_2a_2 + x_3a_3$ 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

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

$$\begin{aligned} f(x) &= \left(f_1\left(\sum_{j=1}^{3} x_j e_j\right), f_2\left(\sum_{j=1}^{3} x_j e_j\right), f_3\left(\sum_{j=1}^{3} x_j e_j\right)\right)^\top \\ &= \left(\sum_{j=1}^{3} f_1(e_j)x_j, \sum_{j=1}^{3} f_2(e_j)x_j, \sum_{j=1}^{3} f_3(e_j)x_j\right)^\top \\ &= \left(\sum_{j=1}^{3} a_{1j}x_j, \sum_{j=1}^{3} a_{2j}x_j, \sum_{j=1}^{3} a_{3j}x_j\right)^\top = Ax, \end{aligned}$$

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

$$\hat{x} \cdot \hat{y} = \left(Q^\top x\right) \cdot \left(Q^\top y\right) = x \cdot \left(Q^\top\right)^\top Q^\top y = x \cdot y,$$

where we used that for any 3×3 matrix $A = (a_{ij})$ and $x, y \in \mathbb{R}^3$

$$\begin{aligned}(Ax) \cdot y &= \sum_{i=1}^{3}\left(\sum_{j=1}^{3} a_{ij}x_j\right) y_i = \sum_{j=1}^{3}\left(\sum_{i=1}^{3} a_{ij}y_i\right) x_j \\ &= \left(A^\top y\right) \cdot x = x \cdot \left(A^\top y\right),\end{aligned} \tag{21.45}$$

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,

$$f(x) = f(x_1, x_2, x_3) = \left(x_2x_3, x_1^2 + x_3, x_3^4 + 5\right)$$

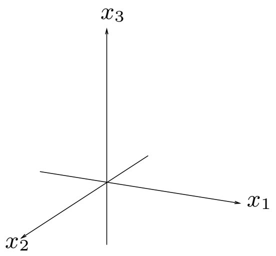
with $f_1(x) = x_2x_3$, $f_2(x) = x_1^2+x_3$, $f_3(x) = x_3^4+5$. We shall see that we may naturally extend the concepts of Lipschitz continuity and differentiability for functions $f : \mathbb{R} \to \mathbb{R}$ to functions $f : \mathbb{R}^3 \to \mathbb{R}^3$. For example, we say that $f : \mathbb{R}^3 \to \mathbb{R}^3$ is Lipschitz continuous on $\mathbb{R}^3$ if there is a constant L_f such that

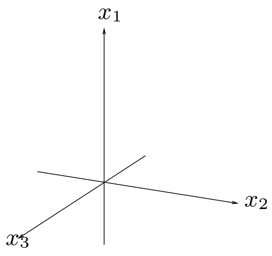
$$|f(x) - f(y)| \leq L_f|x - y| \quad \text{for all } x, y \in \mathbb{R}^3.$$

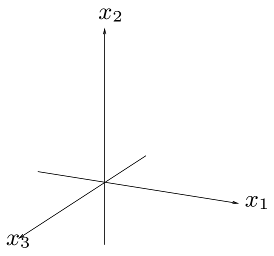
Chapter 21 Problems

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 $\hat{p}_2$.

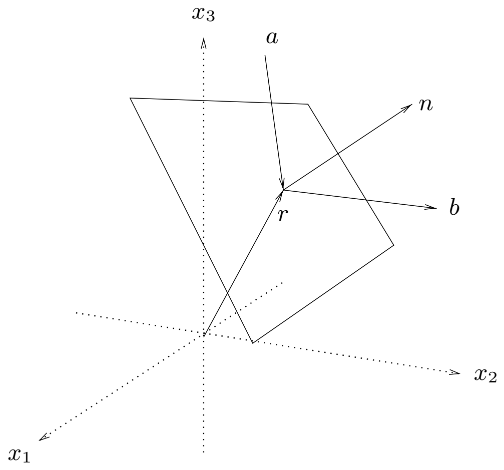
21.17. Consider two straight lines in $\mathbb{R}^3$ given by: $a + \lambda b$ and $c + \mu d$ where $a, b, c, d \in \mathbb{R}^3$, $\lambda, \mu \in \mathbb{R}$. What is the shortest distance between the two lines?

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.

21.34 Gösta Mittag-LefflerTS$^{\text{f}}$

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. 21.34. 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.

Fig. 21.11. 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"

TS$^{\text{f}}$ Please confirm position of Sect. 21.34 only after Problems.

Editor's or typesetter's annotations (will be removed before the final TEX run)

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*.

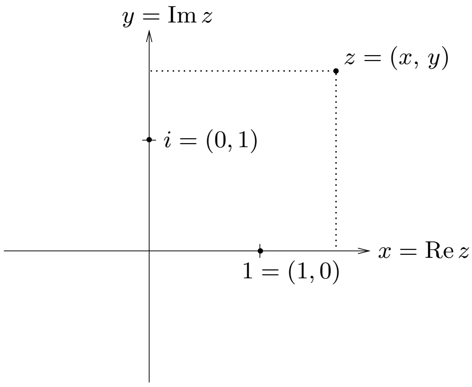


Fig. 22.1. The complex plane $\mathbb{C} = \mathbb{R}^2$

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

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 $x=\mathrm{Re}\ z$ and $y=\mathrm{Im}\ z$, that is

$$z=\mathrm{Re}\ z+i\,\mathrm{Im}\ z=(\mathrm{Re}\ z,\mathrm{Im}\ z). \tag{22.4}$$

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

$$|x|=|\mathrm{Re}\ z|\leq|z|,\qquad |y|=|\mathrm{Im}\ z|\leq|z|. \tag{22.7}$$

22.3 The Triangle Inequality

If z_1 and z_2 are two complex numbers, then

$$|z_1+z_2|\leq|z_1|+|z_2|. \tag{22.8}$$

This is the *triangle inequality for complex numbers*, which follows directly from the triangle inequality in $\mathbb{R}^2$.

22.4 Open Domains

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

$$z = (x, y) = r(\cos(\theta), \sin(\theta)) = r(\cos(\theta) + i\sin(\theta)), \tag{22.9}$$

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 $n = \pm 1, \pm 2, \ldots,$. Choosing $\theta \in [0, 2\pi)$, we obtain the *principal argument* of z, which we denote by $\operatorname{Arg} z$.

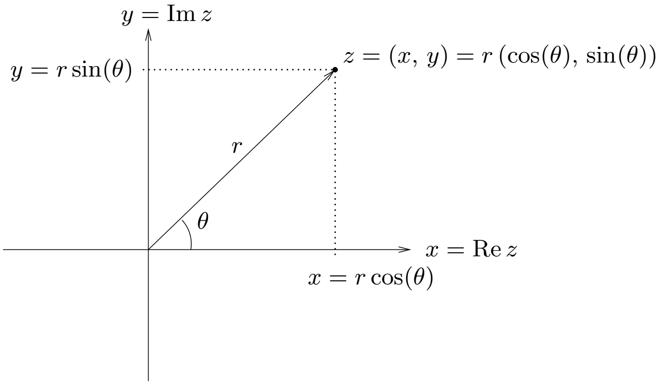

$y = \operatorname{Im} z$
$y = r\sin(\theta)$
$z = (x,\, y) = r\,(\cos(\theta),\, \sin(\theta))$
r
θ
$x = \operatorname{Re} z$
$x = r\cos(\theta)$

Fig. 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

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

$$\arg\, z\zeta = \text{Arg } z + \text{Arg } \zeta, \quad |z\zeta| = |z||\zeta|. \tag{22.10}$$

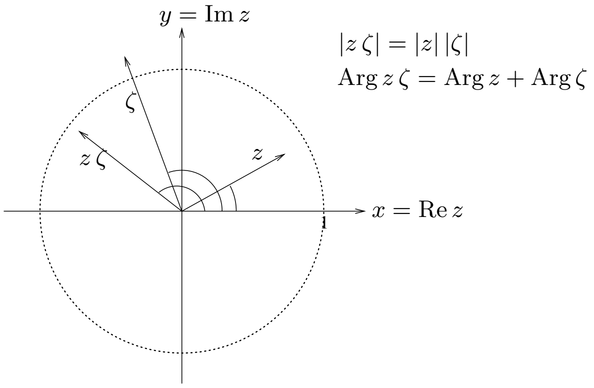

$y = \operatorname{Im} z$
$|z\,\zeta| = |z|\,|\zeta|$
$\operatorname{Arg} z\,\zeta = \operatorname{Arg} z + \operatorname{Arg} \zeta$
ζ
$z\,\zeta$
z
$x = \operatorname{Re} z$
1

Fig. 22.3. Geometrical interpretation of multiplication of a complex numbers

Example 22.2. Multiplication by i corresponds to rotation counter-clockwise $\frac{\pi}{2}$, 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 $\text{Re z} = 0$, 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

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 $p(z) = a_0 + a_1 z + \ldots + a_n z^n$ is a polynomial in z of degree n with complex coefficients $a_0, \ldots, a_n$. 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

$$p(z) = c(z - z_1) \ldots (z - z_n),$$

where $z_1, \ldots, z_n$ 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 $n = 1, 2, \dots$ 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 $k = 0, \dots, n-1$. 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.*

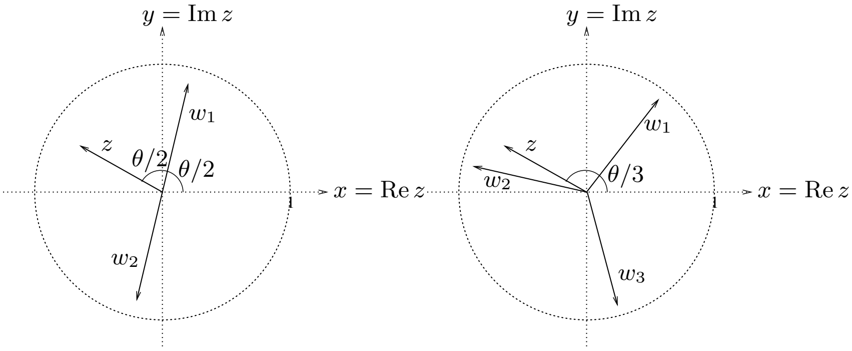


Fig. 22.4. The "square" and "cubic" roots of z

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

$$w = -b \pm i\sqrt{c - b^2}.$$

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) Im $z^2 = 2$, (c) $|\text{Arg } z| \leq \frac{\pi}{4}$.

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:

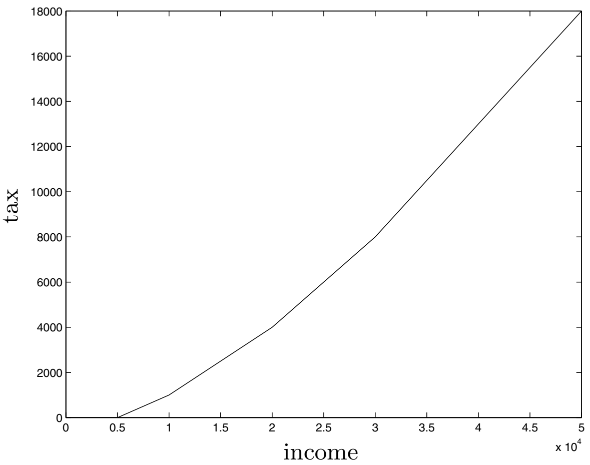


Fig. 23.1. Income tax $f(x)$ varying with income x

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 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.

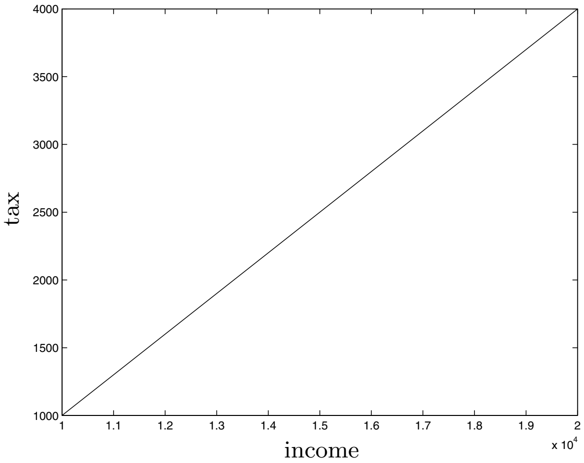


Fig. 23.2. Income tax $f(x)$ for income x in the interval $[10\,000, 20\,000]$

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 $f : \mathbb{R} \rightarrow \mathbb{R}$ 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.

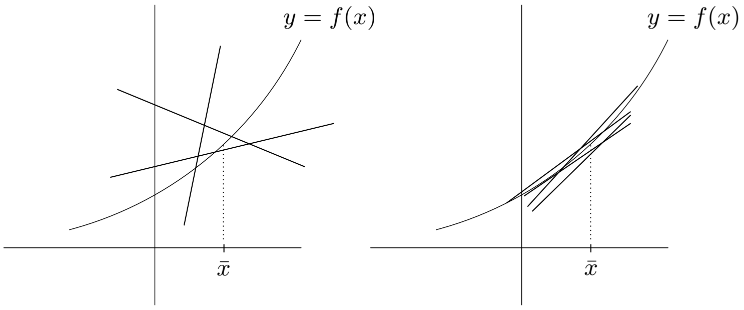


Fig. 23.3. Linear approximations of $f(x)$ close to $\bar{x}$

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.

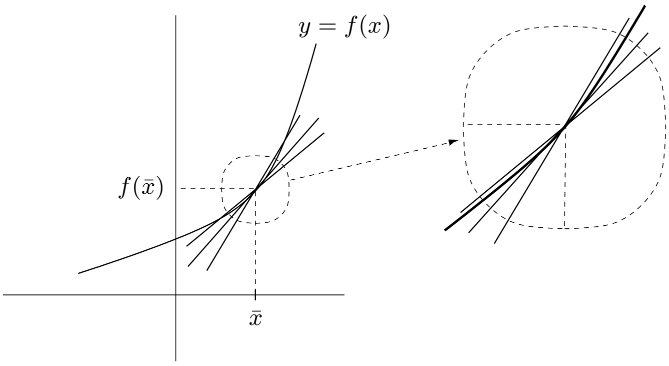


Fig. 23.4. Linear approximations to a function that pass through the point $(\bar{x}, f(\bar{x}))$. The region near $(\bar{x}, f(\bar{x}))$ has been blown-up on the *right*

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 $\bar{x}$.

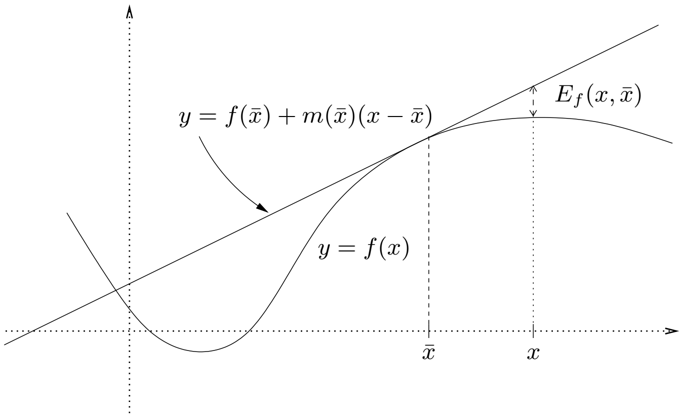


Fig. 23.5. Graph $y = f(x)$, tangent $y = f(\bar{x}) + m(\bar{x})(x-\bar{x})$ and error $E_f(x,\bar{x})$

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 $\bar{x}$,

$$f(x) = f(\bar{x}) + m(\bar{x})(x - \bar{x}) + E_f(x, \bar{x}), \qquad \text{with } |E_f(x, \bar{x})| \leq K_f(\bar{x})|x - \bar{x}|^2. \quad (23.7)$$

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, shorter route to the linearization formula (23.6) in this case is

$$x^2 = (\bar{x} + (x - \bar{x}))^2 = \bar{x}^2 + 2\bar{x}(x - \bar{x}) + (x - \bar{x})^2,$$

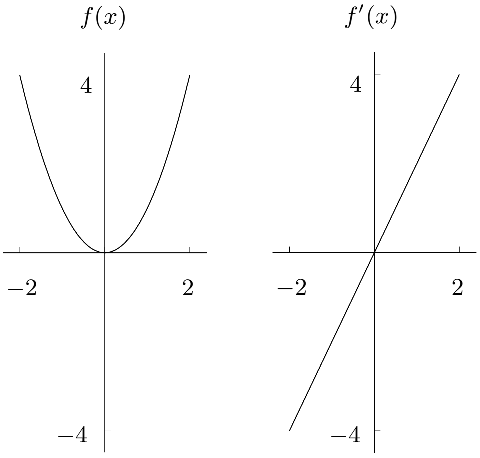


Fig. 23.6. $f(x) = x^2$ and $f'(x) = 2x$

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 to the linear approximation $1 + 2(x - 1)$ along with the error $(x - 1)^2$ in Fig. 23.8

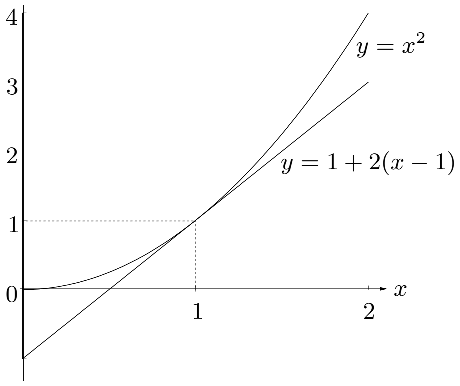


Fig. 23.7. The linearization $1 + 2(x-1)$ of x^2 at $\bar{x} = 1$

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

Fig. 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 $\bar{x} \neq 0$,

$$\frac{1}{x} = \frac{1}{\bar{x}} + \left(\frac{1}{x} - \frac{1}{\bar{x}}\right) = \frac{1}{\bar{x}} + \left(-\frac{1}{x\bar{x}}\right)(x - \bar{x}) = \frac{1}{\bar{x}} + \left(-\frac{1}{\bar{x}^2}\right)(x - \bar{x}) + E$$

where

$$E = \left(\frac{1}{\bar{x}^2} - \frac{1}{x\bar{x}}\right)(x - \bar{x}) = \frac{1}{x\bar{x}^2}(x - \bar{x})^2,$$

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 as a Function

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 $3x^2$.

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

$$f'(x) = Df(x).$$

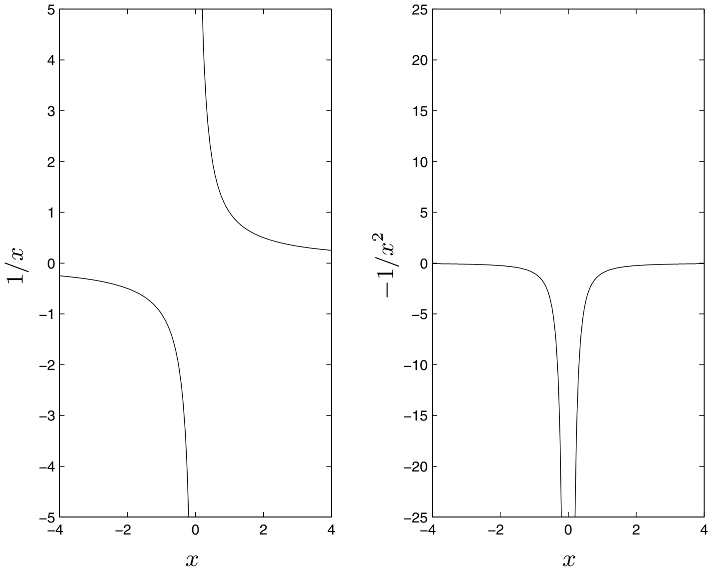


Fig. 23.9. The function $f(x) = 1/x$ and its derivative $f'(x) = -1/x^2$ for $x > 0$

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

$$Dx^n = nx^{n-1} \quad \text{for } n = 1, 2, \ldots \tag{23.13}$$

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.

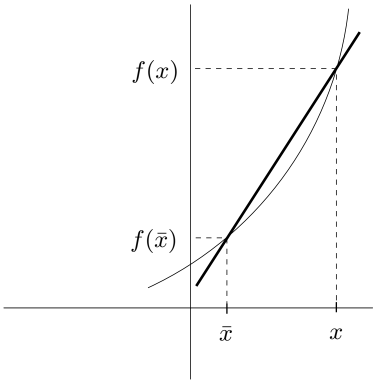


Fig. 23.10. The secant line joining $(\bar{x}, f(\bar{x}))$ and $(x_i, f(x_i))$

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.

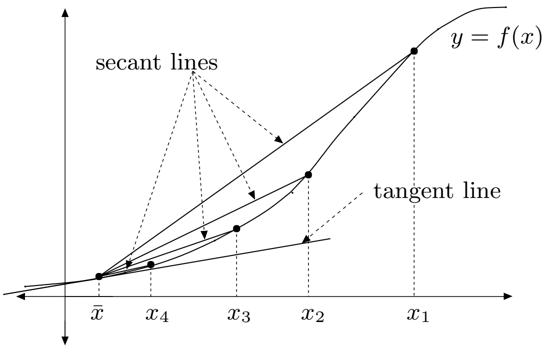


Fig. 23.11. A sequence of secant lines approaching the tangent line at $\bar{x}$

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

$$f'(\bar{x}) = \lim_{i\to\infty} m_i(\bar{x}) = \lim_{i\to\infty} \left(2\bar{x} + \frac{1}{i}\right) = 2\bar{x},$$

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

$$\left|\frac{E_f(x,\bar{x})}{x-\bar{x}}\right|\leq K_f(\bar{x})|x-\bar{x}|.$$

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

$$\left|\frac{\delta f}{x-\bar{x}}\right|+K_f(\bar{x})|x-\bar{x}|, \tag{23.25}$$

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

$$\left|\frac{\delta f}{x-\bar{x}}\right|=K_f|x-\bar{x}|,$$

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

$$\left| f'(\bar{x}) - \frac{f(x) - f(\bar{x})}{x - \bar{x}} \right| \le 2\sqrt{\delta f}\sqrt{K_f}. \tag{23.27}$$

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 $\bar{x}$,

$$\begin{aligned} f(x) &= (f(\bar{x}) + m(\bar{x})(x - \bar{x})) + E_f(x, \bar{x}) \\ |E_f(x, \bar{x})| &\le K_f(\bar{x})|x - \bar{x}|^2. \end{aligned}$$

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 $f : I \rightarrow \mathbb{R}$ 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 A 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.

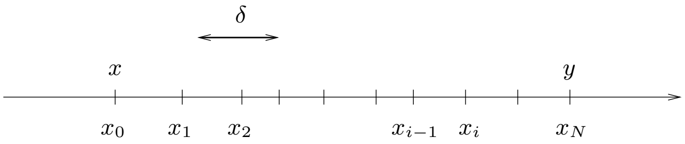


Fig. 23.12. Subdivision of interval $[x, y]$ into subintervals of length $< \delta$

We have by the triangle inequality

$$\begin{aligned} |f(x) - f(y)| &= \left| \sum_{i=1}^{N} (\text{TS}^{g} f(x_i) - f(x_{i-1}) \right| \\ \leq \sum_{i=1}^{N} |f(x_i) - f(x_{i-1})| &\leq (L + K\delta) \sum_{i=1}^{N} |x_i - x_{i-1}| \\ &= (L + K\delta)|x - y|. \end{aligned}$$

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.

TSg Please check this opening parenthesis.

Editor's or typesetter's annotations (will be removed before the final TEX run)

23.16 A Slightly Different Viewpoint

In many Calculus books the derivative of a function $f : \mathbb{R} \to \mathbb{R}$ at a point $\bar{x}$ is defined as follows. If the limit

$$\lim_{i\to\infty} \frac{f(x_i) - f(\bar{x})}{x_i - \bar{x}}, \tag{23.30}$$

does exist for any sequence $\{x_i\}$ with $\lim_{i\to\infty} x_i = \bar{x}$ (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

$$f'(\bar{x}) = \lim_{i\to\infty} \frac{f(x_i) - f(\bar{x})}{x_i - \bar{x}},$$

because we assume that

$$\left| f'(\bar{x}) - \frac{f(x_i) - f(\bar{x})}{x_i - \bar{x}} \right| \leq K_f(\bar{x})|x_i - \bar{x}|. \tag{23.31}$$

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

Fig. 23.13. Emanuel Swedenborg, Swedish Universal Genius, as a young man: "The Intercourse of Soul and Body is thus not effected by any physica influx or by any action of the Body upon the Mind or Soul; for the lower cannot affect the higher, and the nature cannot inflow into the spiritual. Yet the Soul can accomodate itself to the changes of the sensories of the brain and form mental percepts and concepts. It can also time the release of the energy there stored and from an intelligent conatus direct it into motivated or living actions"

of the (at the time) world's largest dry-dock, and suggested the nebula theory of the formation of the planets.

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

$$f'(\bar{x}) \approx \frac{f(\bar{x}+h)-f(\bar{x}-h)}{h} \qquad h > 0.$$

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

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n = \sum_{i=0}^{n} a_i x^i$$

is

$$f'(x) = a_1 + 2a_2 x^2 + \cdots + na_n x^{n-1} = \sum_{i=1}^{n} ia_i x^{i-1}.$$

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.

$$\begin{aligned}D\,\big((10 + 3x^2 - x^6)(x - 7x^4)\big) \\ = (6x - 6x^5)(x - 7x^4) + (10 + 3x^2 - x^6)(1 - 28x^3).\end{aligned}$$

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

$$|E_f(y, \bar{y})| \le K_f |y - \bar{y}|^2, \quad |E_g(x, \bar{x})| \le K_g |x - \bar{x}|^2, \tag{24.12}$$

where K_f and K_g are certain constants, independent of y and x, respectively. Further, by assumption

$$|g(x) - g(\bar{x})| \le L_g |x - \bar{x}|. \tag{24.13}$$

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

$$\begin{aligned} f(g(x)) &= f(g(\bar{x})) + f'(g(\bar{x}))\, g'(\bar{x})(x - \bar{x}) \\ &\quad + f'(g(\bar{x})) E_g(x, \bar{x}) + E_f(g(x), g(\bar{x})). \end{aligned}$$

Since (24.12) and (24.13) imply

$$\begin{aligned} |E_f(g(x), g(\bar{x}))| &\le K_f |g(x) - g(\bar{x})|^2 \le K_f L_g^2 |x - \bar{x}|^2, \\ |f'(g(\bar{x}))\, E_g(x, \bar{x})| &\le |f'(g(\bar{x}))|\, K_g |x - \bar{x}|^2, \end{aligned}$$

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 $m = -1, -2, \ldots$

24.5 The Quotient Rule

Let $f(x)$ and $g(x)$ be differentiable on I and consider the problem of computing the derivative 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:

$$\begin{aligned}
&\frac{d}{dx}\left(\left(\left((1-x)^2+1\right)^3+2\right)^4+3\right)^5 \\
&\quad = 5\left(\left(\left((1-x)^2+1\right)^3+2\right)^4+3\right)^4 \times 4\left(\left((1-x)^2+1\right)^3+2\right)^3 \\
&\qquad\qquad \times 3\left((1-x)^2+1\right)^2 \times 2(1-x) \times (-1).
\end{aligned}$$

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 $x \in [\bar{x}, b)$

$$|f(x) - (f(\bar{x}) + m(\bar{x})(x - \bar{x}))| \leq K_f(\bar{x})|x - \bar{x}|^2. \tag{24.18}$$

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 $f : [a, b] \rightarrow \mathbb{R}$ 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

$$f(x) = f(\bar{x}) + m_1(\bar{x})(x - \bar{x}) + m_2(\bar{x})(x - \bar{x})^2 + E_f(x, \bar{x}), \qquad (24.19)$$

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

$$|E_f(x, \bar{x})| \leq K_f(\bar{x})|x - \bar{x}|^3, \qquad (24.20)$$

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

$$f(x) = f(\bar{x}) + f'(\bar{x})(x-\bar{x}) + m_2(\bar{x})(x-\bar{x})^2 + E_f(x,\bar{x}). \qquad (24.21)$$

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

$$f'(x) = f'(\bar{x}) + 2m_2(\bar{x})(x-\bar{x}) + \frac{d}{dx}E_f(x,\bar{x}). \qquad (24.22)$$

Let us now assume that for x close to $\bar{x}$

$$\left|\frac{d}{dx}E_f(x,\bar{x})\right| \leq M_f(\bar{x})|x-\bar{x}|^2, \qquad (24.23)$$

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

$$f(x) = f(\bar{x}) + f'(\bar{x})(x-\bar{x}) + \frac{1}{2}f''(\bar{x})(x-\bar{x})^2 + E_f(x,\bar{x}), \qquad (24.24)$$

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

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

$$= 1 - y + y^2\frac{1}{1+y} = 1 - y + y^2 - y^3\frac{1}{1+y}.$$

Choosing $y = x-1$, we get

$$\frac{1}{x} = \frac{1}{1+(x-1)} = 1 - (x-1) + (x-1)^2 - \frac{(x-1)^3}{1+(x-1)}, \qquad (24.25)$$

and we see that the quadratic polynomial

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

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.

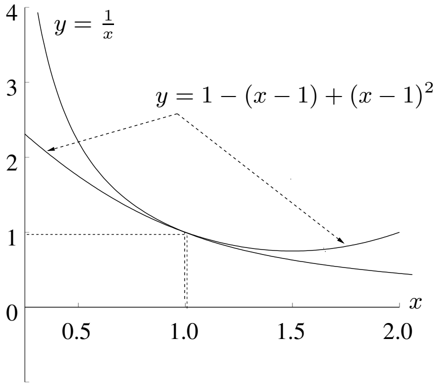


Fig. 24.1. The quadratic approximation $1 - (x-1) + (x-1)^2$ of $1/x$ near $\bar{x} = 1$

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

Fig. 24.2. Some values of $f(x) = 1/x$, the quadratic approximation $1 - (x-1) + (x-1)^2$, and the error $E_f(x,1)$

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 Fig. 24.3.

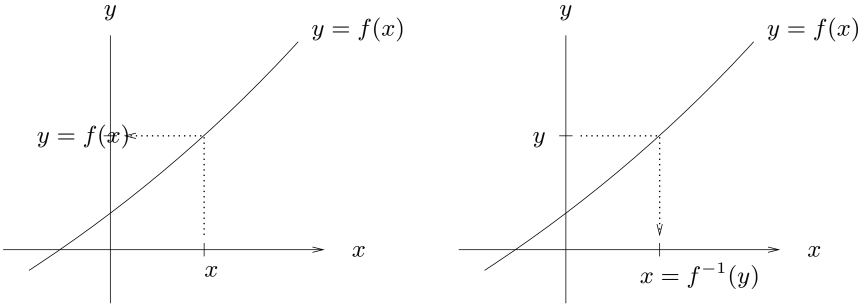


Fig. 24.3. The function $y=f(x)$ and its inverse $x=f^{-1}(y)$

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 $\bar{y}$,

$$|\frac{1}{f'(\bar{x})}\,E_f(f^{-1}(y),f^{-1}(\bar{y}))|\leq\frac{1}{|f'(\bar{x})|}K_f(\bar{x})(L_{f^{-1}})^2|y-\bar{y}|^2,$$

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 $f(x_1, x_2) = x_1^2 + x_2^3 + x_1x_2$. 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

$$|E_f(x, \bar{x})| \leq K_f(\bar{x}_1, \bar{x}_2)((x_1 - \bar{x}_1)^2 + (x_2 - \bar{x}_2)^2).$$

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 $f(x_1, \ldots, x_d)$ of d real variables $x_1, \ldots, x_d$, and we can speak about (and compute) partial derivatives of $f(x_1, \ldots, x_d)$ with respect to $x_1, \ldots, x_d$ following the same basic idea. To compute the partial derivative $\frac{\partial f}{\partial x_j}$ with respect to x_j for some $j = 1, \ldots, d$, 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

$$\begin{aligned} Dx^n &= \frac{d}{dx}x^n = nx^{n-1} \quad \text{for } n \text{ integer and } x \neq 0, \\ Dx^r &= \frac{d}{dx}x^r = rx^{r-1} \quad \text{for } r \text{ rational and } x > 0. \end{aligned}$$

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 $i = 1, 2, \ldots$. 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 $|g'(\bar{x})|$ 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 $i = 0, 1, 2, \ldots$

$$x_{i+1} = x_i - \frac{f(x_i)}{f'(x_i)}, \tag{25.5}$$

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 $[-3, 3]$ and indicate the corresponding roots that are found 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

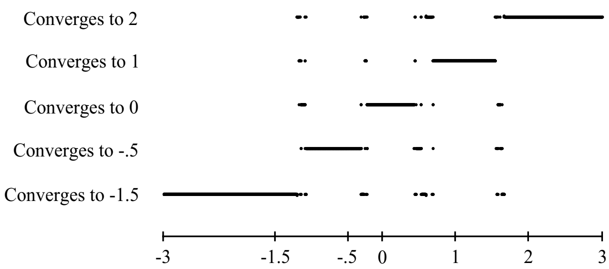


Fig. 25.1. This plot shows the roots of $f(x) = (x-2)(x-1)x(x+.5)(x+1.5)$ found by Newton's method for 5000 equally spaced initial guesses in $[-3,3]$. The horizontal position of the points shows the location of the initial guess and the vertical position indicates the twenty first Newton iterate

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 $f : \mathbb{R}^n \to \mathbb{R}^n$.

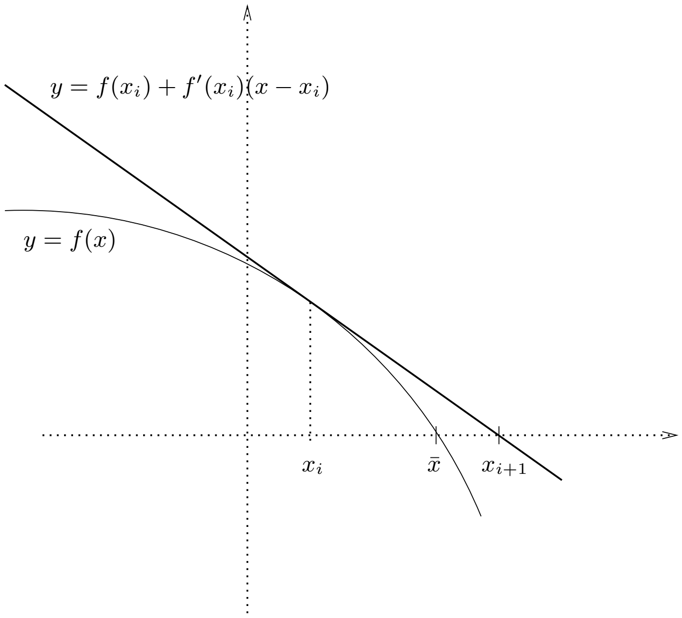


Fig. 25.2. An illustration of one step of Newton's method from x_i to x_{i+1}

25.6 What Is the Error of 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. 25.3.

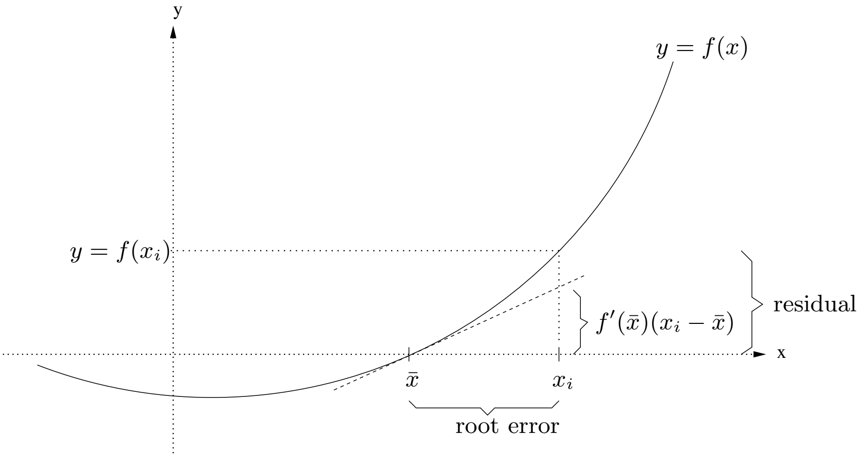


Fig. 25.3. The root error and the residual

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 $f'(\bar{x}) \approx 0.0000000316$,

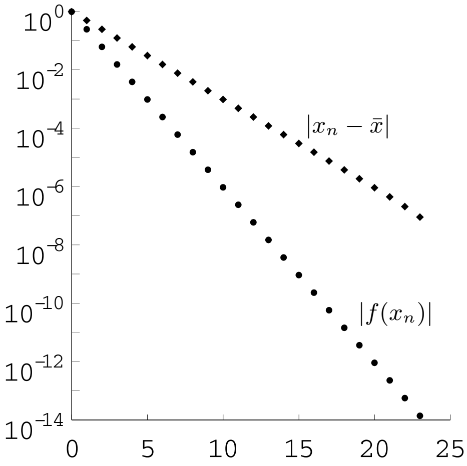


Fig. 25.4. Plots of the residuals ● and errors ⧫ versus iteration number for Newton's method applied to $f(x) = (x-1)^2 - 10^{-15}x$ with initial value $x_0 = 1$

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. 25.4. 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 explicitly.

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

Fig. 25.5. The error and error estimate for Newton's method for $f(x) = x^2 - 2$ with $x_0 = 2$

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. 25.5. The error estimate does a pretty good job.

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

$$|x_i - \bar{x}| \leq TOL, \tag{25.10}$$

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, 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.

Fig. 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-

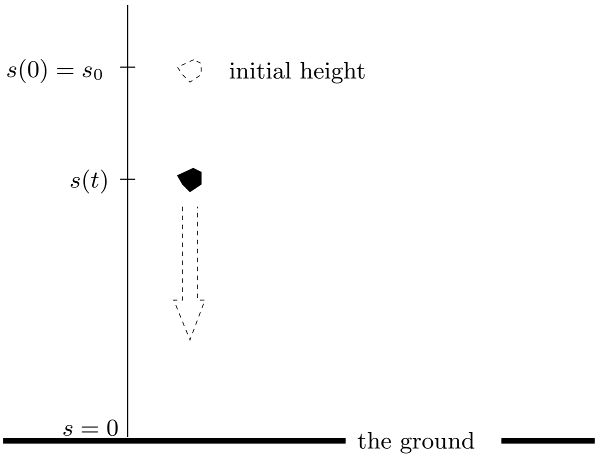


Fig. 26.2. The coordinate system describing the position of a free falling object with the initial height $s(0) = s_0$ at time $t = 0$

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 un-stretched 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 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 $\omega = \sqrt{\frac{k}{m}}$.

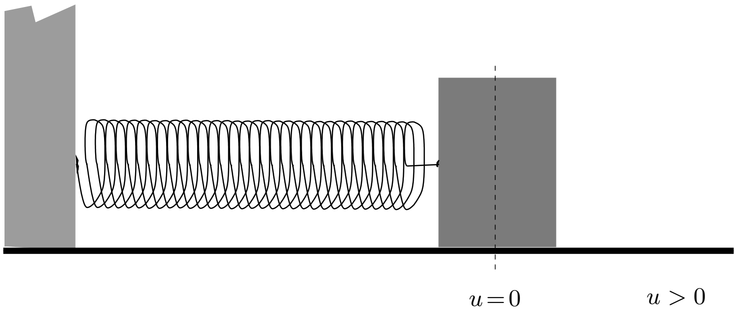


Fig. 26.3. Illustration of the coordinate system used to describe a spring-mass system. The mass is allowed to slide freely back and forth with no friction

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 $n = 1, 2, \ldots$, 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 $t = 0$.

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

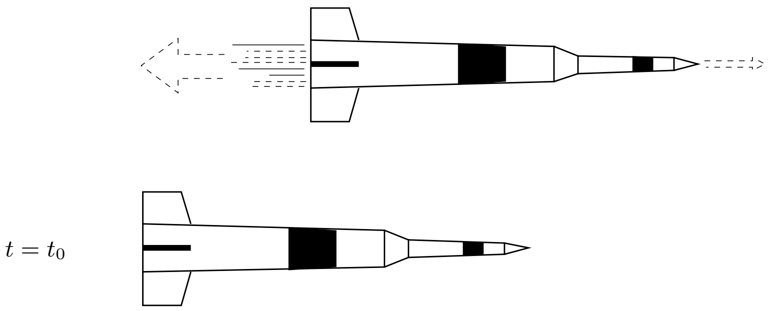


Fig. 26.4. A model of rocket propulsion

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 and $\Delta m_e(v(t) - u)$ the momentum of the exhaust released over $(t, t+\Delta t)$. 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).

Fig. 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 $t_n = kn$, $n = 0, 1, 2, \ldots$ 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

$$U_n = U_{n-1} + k\lambda U_{n-1}, \quad \text{for } n = 1, 2, 3 \ldots, \quad U_0 = 1. \tag{26.18}$$

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

$$U_n = (1 + k\lambda)^n, \quad \text{for } n = 0, 1, 2, \dots \tag{26.19}$$

We expect thus that

$$u(kn) \approx (1 + k\lambda)^n, \quad \text{for } n = 0, 1, 2, \dots \tag{26.20}$$

or

$$u(t) \approx \left(1 + \frac{t}{n}\lambda\right)^n, \quad \text{for } n = 0, 1, 2, \dots \tag{26.21}$$

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

$$\exp(\lambda t) = e^{\lambda t} = \lim_{n\to\infty}\left(1 + \frac{\lambda t}{n}\right)^n, \tag{26.23}$$

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

$$v^2(t) = c^2 \, \frac{F^2t^2}{m_0^2c^2 + F^2t^2}. \tag{26.26}$$

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

$$s'(t) = v(t) = \frac{cFt}{\sqrt{m_0^2c^2 + F^2t^2}}. \tag{26.27}$$

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.

> Show that the median, hce che ech, interecting at royde angles the parilegs of a given obtuse one biscuts both the arcs that are in curveachord behind. Brickbaths. The family umbroglia. A Tullagrove pole to the Heigt of County Fearmanagh has a septain inclininaison and the graphplot for all the functions in Lower County Monachan, whereat samething is rivisible by nighttim. may be involted into the zerois couplet, palls pell inhis heventh glike noughty times ∞, find, if you are literally cooefficient, how minney combinaisies and permutandis can be played on the international surd! pthwndxrclzp!, hids cubid rute being extructed, takin anan illitterettes, ifif at a tom. Answers, (for teasers only). (Finnegans Wake, James Joyce)

References

[1] L. Ahlfors, *Complex Analysis*, McGraw-Hill Book Company, New York, 1979.

[2] K. Atkinson, *An Introduction to Numerical Analysis*, John Wiley and Sons, New York, 1989.

[3] L. Bers, *Calculus*, Holt, Rinehart, and Winston, New York, 1976.

[4] M. Braun, *Differential Equations and their Applications*, Springer-Verlag, New York, 1984.

[5] R. Cooke, *The History of Mathematics. A Brief Course*, John Wiley and Sons, New York, 1997.

[6] R. Courant and F. John, *Introduction to Calculus and Analysis*, vol. 1, Springer-Verlag, New York, 1989.

[7] R. Courant and H. Robbins, *What is Mathematics?*, Oxford University Press, New York, 1969.

[8] P. Davis and R. Hersh, *The Mathematical Experience*, Houghton Mifflin, New York, 1998.

[9] J. Dennis and R. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, New Jersey, 1983.

[10] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Computational Differential Equations*, Cambridge University Press, New York, 1996.

[11] I. Grattan-Guiness, *The Norton History of the Mathematical Sciences*, W.W. Norton and Company, New York, 1997.

[12] P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley and Sons, New York, 1962.

[13] E. Isaacson and H. Keller, *Analysis of Numerical Methods*, John Wiley and Sons, New York, 1966.

[14] M. Kline, *Mathematical Thought from Ancient to Modern Times*, vol. I, II, III, Oxford University Press, New York, 1972.

[15] J. O'Connor and E. Robertson, *The MacTutor History of Mathematics Archive*, School of Mathematics and Statistics, University of Saint Andrews, Scotland, 2001. `http://www-groups.dcs.st-and.ac.uk/~history/`.

[16] W. Rudin, *Principles of Mathematical Analysis*, McGraw–Hill Book Company, New York, 1976.

[17] T. Ypma, *Historical development of the Newton-Raphson method*, SIAM Review, 37 (1995), pp. 531–551.

Index